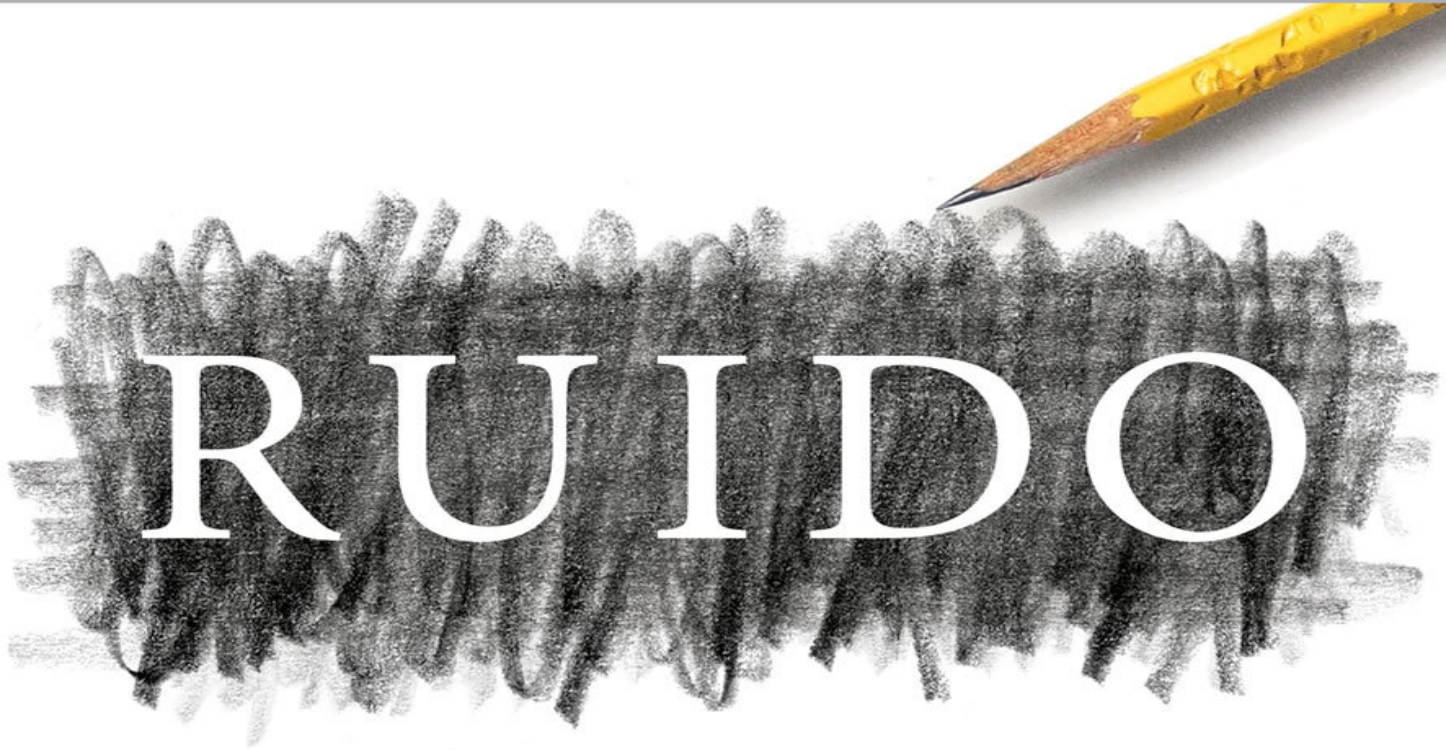

Un fallo en el juicio humano

DANIEL KAHNEMAN

AUTOR DE *PENSAR RÁPIDO, PENSAR DESPACIO*

OLIVIER SIBONY

CASS R. SUNSTEIN

DEBATE

Ruido

Un fallo en el juicio humano

DANIEL KAHNEMAN
OLIVIER SIBONY
CASS R. SUNSTEIN

Traducción de
Joaquín Chamorro Mielke

DEBATE

A Noga, Ori y Gili (DK)
A Fantin y Lélia (OS)
A Samantha (CRS)

Introducción
Dos clases de error

Imaginemos que cuatro equipos de amigos han acudido a un campo de tiro. Cada equipo está formado por cinco personas; comparten un rifle y cada una realiza un disparo. La figura 1 muestra sus resultados. En un mundo ideal, todos los disparos darían en la diana.

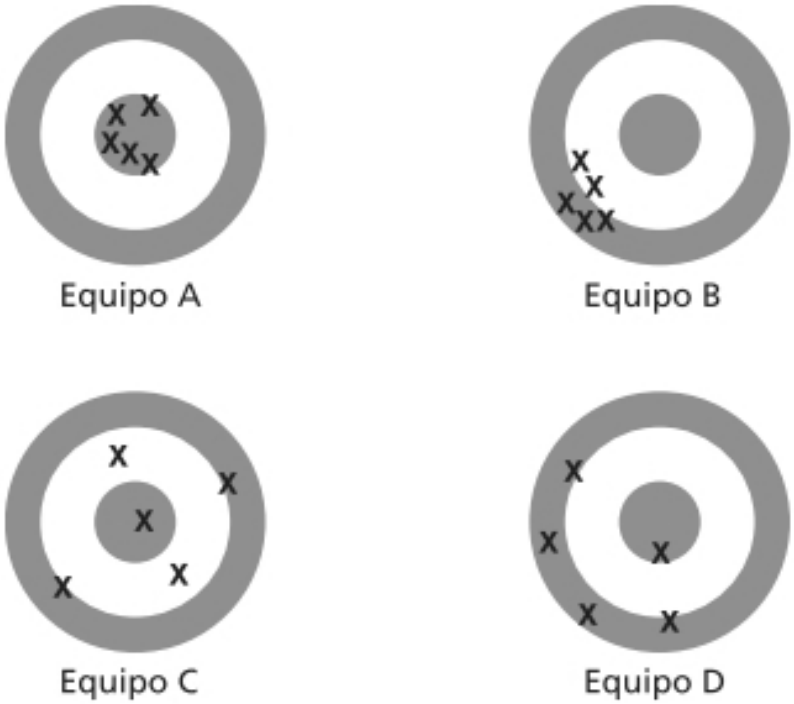

FIGURA 1. Cuatro equipos

Casi es lo que ocurre con el equipo A. Los disparos de este equipo están muy agrupados alrededor de la diana, cerca de un patrón perfecto.

De los resultados del equipo B diríamos que están «sesgados», porque sus disparos están sistemáticamente fuera del objetivo. Como ilustra la figura 1, la consistencia del sesgo apoya una predicción. Si uno de los miembros del equipo hiciera otro disparo, apostaríamos por un impacto en la misma zona que los cinco primeros. La consistencia del sesgo también

invita a una explicación causal: quizá la mira del rifle del equipo estaba doblada.

Del equipo C diríamos que es «ruidoso», porque sus disparos se hallan muy dispersos. No hay un sesgo evidente, porque los impactos están centrados aproximadamente en la diana. Si uno de los miembros del equipo realizara otro disparo, sabríamos muy poco sobre dónde es probable que impacte. Además, no se nos ocurre ninguna hipótesis interesante para explicar los resultados del equipo C. Sabemos que sus miembros son malos tiradores. Desconocemos por qué son tan ruidosos.

El equipo D está sesgado y es ruidoso. Como el equipo B, sus disparos se desvían sistemáticamente del objetivo y, al igual que el equipo C, estos están muy dispersos.

Sin embargo, este no es un libro sobre tiro al blanco. Nuestro tema es el error humano. El sesgo y el ruido —desviación sistemática y dispersión aleatoria— son componentes diferentes del error. Los blancos ilustran la diferencia.[1]

El campo de tiro es una metáfora de lo que puede fallar en el juicio humano, en particular en las diversas decisiones que se toman en nombre de organizaciones. En estas situaciones, nos encontraremos con las dos clases de error que ilustra la figura 1. Unos juicios están sesgados; se desvían sistemáticamente del objetivo. Otros son ruidosos, ya que las personas que se espera que estén de acuerdo terminan en puntos muy diferentes alrededor del objetivo. Por desgracia, muchas organizaciones se ven afectadas tanto por el sesgo como por el ruido.

La figura 2 ilustra una importante diferencia entre el sesgo y el ruido. Muestra lo que se vería en el campo de tiro si solo se mostrara el dorso de los blancos a los que disparaban los equipos (sin nada que indique la posición de la diana a la que apuntaban).

El dorso de los blancos no nos permite saber si el equipo A o el equipo B está más cerca de la diana; pero a simple vista se puede apreciar que los equipos C y D son ruidosos y que los equipos A y B no lo son. Aquí sabemos tanto de la dispersión como en la figura 1. Una propiedad general del ruido es que se puede reconocer y medir sin saber nada sobre el objetivo o el sesgo.

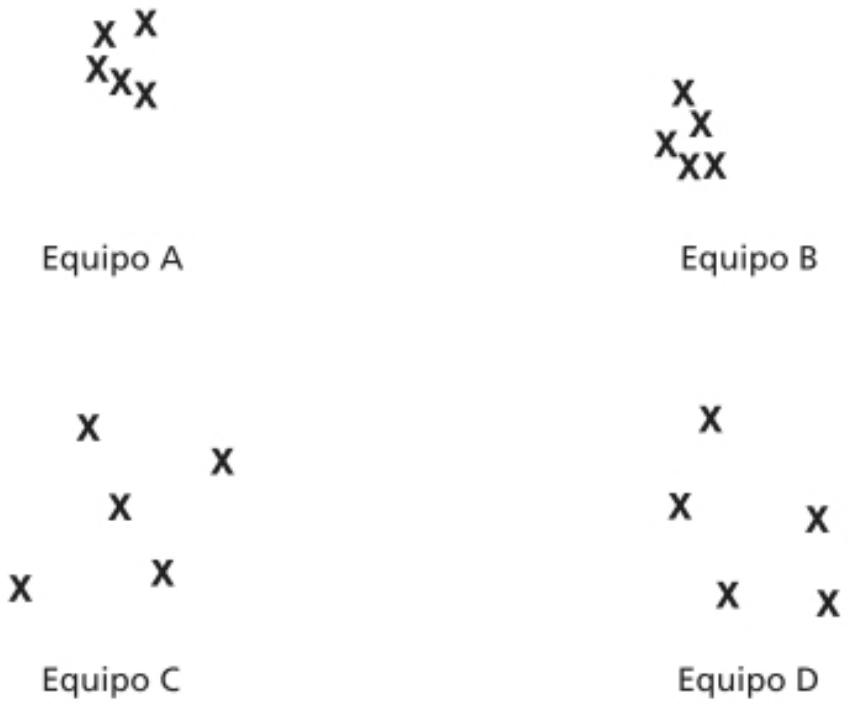

FIGURA 2. Aspecto del dorso de las dianas

La propiedad general del ruido que acabamos de mencionar es esencial para nuestros propósitos en este libro, porque muchas de nuestras conclusiones se extraen de juicios cuyo objetivo real es desconocido o incluso imposible de conocer. Cuando los médicos ofrecen diferentes diagnósticos para el mismo paciente, podemos estudiar su desacuerdo sin saber qué le ocurre al paciente. Cuando los ejecutivos del cine estiman las posibilidades que tendrá una película en el mercado, podemos estudiar la variabilidad de sus respuestas sin saber lo que consiguió finalmente la película, o incluso si se llegó a filmar. No necesitamos saber quién tiene razón para medir cuánto varían los juicios de un mismo caso. Todo lo que tenemos que hacer para medir el ruido es mirar la parte trasera del objetivo.

Para comprender el error en el juicio, necesitamos entender tanto el sesgo como el ruido. A veces el ruido es, como veremos, el problema más importante. Sin embargo, en las conversaciones públicas sobre el error humano y en las organizaciones de todo el mundo, el ruido rara vez se reconoce. El sesgo es la estrella del espectáculo. El ruido es un actor de reparto que normalmente está fuera del escenario. El tema del sesgo ha sido

discutido en miles de artículos científicos y en decenas de libros populares, pocos de los cuales mencionan el asunto del ruido. Este libro intenta restablecer el equilibrio.

En las decisiones del mundo real, la cantidad de ruido suele ser altísima. He aquí algunos ejemplos de la alarmante cantidad de ruido que se da en situaciones en que la exactitud es importante:

- *La medicina es ruidosa*. Ante un mismo paciente, distintos médicos juzgan de forma diferente si este tiene cáncer de piel, cáncer de mama, enfermedad cardiaca, tuberculosis, neumonía, depresión y un sinfín de afecciones. El ruido es muy alto en psiquiatría, donde el juicio subjetivo es importante. Sin embargo, también se encuentra un ruido considerable en ámbitos donde no cabría esperarlo, como en la interpretación de radiografías.
- *Las decisiones sobre la custodia infantil son ruidosas*.[2] Los asesores de los organismos de protección de la infancia deben evaluar si los niños corren el riesgo de ser víctimas de maltratos y, en caso afirmativo, colocarlos en un centro de acogida. El sistema es ruidoso, dado que algunos asesores son mucho más propensos que otros a enviar a un niño a un centro de acogida. Años después, la mayoría de los desventurados niños a los que estos asesores de mano dura han asignado un centro de acogida tendrán malos resultados en la vida: mayores tasas de delincuencia, mayores índices de natalidad en la adolescencia y menores ingresos.
- *Las predicciones son ruidosas*. Los analistas profesionales ofrecen predicciones muy variables sobre las probables ventas de un nuevo producto, el probable crecimiento de la tasa de desempleo, la probabilidad de quiebra de empresas con problemas y casi sobre cualquier cosa. No solo están en desacuerdo entre ellos, sino que lo están también consigo mismos. Por ejemplo, cuando se pidió a los mismos programadores informáticos en dos días distintos que calcularan el tiempo de ejecución de la misma tarea, las horas que estimaron difirieron en un 71 por ciento de promedio. [3]
- *Las decisiones sobre asilo político son ruidosas*. Que un solicitante de asilo sea admitido en Estados Unidos depende de algo similar a una lotería. Un estudio de casos que fueron asignados al azar a diferentes jueces, encontró que un juez admitió al 5 por ciento de los solicitantes, mientras que otro admitió al 88 por ciento. El título del estudio lo dice todo: «La ruleta de los refugiados». [4] (Vamos a ver muchas ruletas.)
- *Las decisiones en la selección de personal son ruidosas*. Los entrevistadores de los candidatos a un puesto de trabajo hacen evaluaciones muy diferentes de las mismas personas. Las calificaciones del rendimiento de los mismos empleados también son muy variables; dependen más de la persona que realiza la evaluación que del rendimiento que se evalúa.
- *Las decisiones sobre la libertad bajo fianza son ruidosas*. El hecho de que se conceda la libertad bajo fianza a un acusado o, por el contrario, se le envíe a la cárcel en espera de juicio depende en gran medida de la identidad del juez que acabe encargándose del caso. Hay jueces mucho más indulgentes que otros. Los jueces también difieren notablemente en su evaluación del riesgo de fuga o de reincidencia en los acusados.
- *La ciencia forense es ruidosa*. Hemos sido educados para pensar que la identificación de huellas dactilares es infalible. Sin embargo, existe una variabilidad real en los juicios de los dactiloscopistas sobre si una huella encontrada en el lugar del crimen coincide con la de un sospechoso. Los expertos no solo están en desacuerdo; a veces, ellos mismos toman decisiones

inconsistentes cuando se les presenta la misma huella en diferentes ocasiones. Se ha documentado una variabilidad similar en otras disciplinas de la ciencia forense, incluso en análisis de ADN.

- *Las decisiones en la concesión de patentes son ruidosas*. «El que la oficina de patentes conceda o rechace una patente está relacionado de manera significativa con la accidentalidad del examinador que le toque a quien la solicite.»[5] Esta variabilidad es obviamente problemática desde el punto de vista de la equidad.

Todas estas situaciones ruidosas son la punta de un gran iceberg. Dondequiera que examinemos juicios humanos, es probable que encontremos ruido. Para mejorar la calidad de nuestros juicios, necesitamos domeñar tanto el ruido como el sesgo.

Este libro consta de cinco partes. En la primera parte exploramos la diferencia entre el ruido y el sesgo y mostramos que tanto las organizaciones públicas como las privadas pueden ser ruidosas, a veces de una manera asombrosa. Para apreciar el problema, comenzamos con los juicios en dos áreas. La primera, en relación con las sentencias penales (y, por tanto, en el sector público). La segunda, en relación con los seguros (y, por tanto, en el sector privado). A primera vista, las dos áreas no podrían ser más diferentes. Sin embargo, en lo referente al ruido, tienen mucho en común. Para establecer este punto, introducimos la idea de una auditoría del ruido diseñada para medir el desacuerdo que existe entre profesionales que consideran los mismos casos dentro de una organización.

En la segunda parte investigamos la naturaleza del juicio humano y exploramos los modos de medir el acierto y el error. Los juicios son susceptibles de contener tanto sesgos como ruido. Describimos una sorprendente equivalencia en el papel de los dos tipos de error. El ruido de ocasión es la variabilidad en los juicios de un mismo caso hechos por la misma persona o un mismo grupo en ocasiones diferentes. En las discusiones de grupo se da una cantidad sorprendente de ruido de ocasión debido a factores en apariencia irrelevantes, como quién habla primero.

En la tercera parte examinamos más a fondo un tipo de juicio que se ha investigado extensamente, el juicio predictivo. Exploramos la ventaja clave de las reglas, las fórmulas y los algoritmos con respecto a los humanos cuando se trata de hacer predicciones: en contra de la creencia popular, esa ventaja no se debe tanto a la superioridad de las reglas como a su ausencia de ruido. Analizamos el margen último de la calidad del juicio predictivo —la ignorancia objetiva del futuro— y el modo en que conspira con el ruido

para limitar la calidad de la predicción. Finalmente, abordamos una cuestión que es casi seguro que el lector ya se habrá planteado: si el ruido se halla tan extendido, ¿por qué no se había notado antes?

La cuarta parte se vuelve hacia la psicología humana. Explicamos las principales causas del ruido. Entre ellas figuran las diferencias interpersonales creadas por una variedad de factores, como la personalidad y el estilo cognitivo, las variaciones idiosincrásicas en la ponderación de diferentes consideraciones y los distintos usos que las personas hacen de las mismas escalas. Analizamos por qué las personas son ajenas al ruido y a menudo no les sorprenden acontecimientos y juicios que posiblemente no pudieron haber predicho.

La quinta parte examina la cuestión práctica del modo en que podemos mejorar nuestros juicios y evitar el error. (Los lectores que estén interesados sobre todo en las aplicaciones prácticas de la reducción del ruido pueden saltarse el debate sobre los retos de la predicción y la psicología del juicio en las partes tercera y cuarta y pasar directamente a esta parte.) Investigamos los intentos de abordar el ruido en la medicina, los negocios, la educación, el gobierno y otras áreas. Introducimos una serie de técnicas de reducción del ruido con el rótulo de «higiene de las decisiones». Presentamos cinco estudios de casos en campos en los que hay mucho ruido documentado y en los que se han hecho continuos esfuerzos por reducirlo con grados de éxito instructivamente variables. Los estudios de casos incluyen los diagnósticos médicos no fiables, las calificaciones del rendimiento, la ciencia forense, las decisiones de contratación y la predicción en general. Concluimos ofreciendo un sistema que llamamos el «protocolo de evaluaciones mediadoras»: un método de uso general para la evaluación de opciones que incorpora varias prácticas clave de higiene de las decisiones y que tiene por finalidad producir menos ruido y más juicios fiables.

¿Cuál es el nivel de ruido adecuado? La sexta parte trata de esta cuestión. Tal vez parezca ilógico que el nivel correcto no sea cero. En algunas áreas no es factible eliminar el ruido; en otras, es demasiado caro hacerlo. Y, en otras más, los intentos de reducir el ruido comprometerían importantes valores en competencia. Por ejemplo, los intentos de eliminar el ruido pueden socavar la moral y dar a las personas la sensación de que están siendo tratadas como engranajes de una máquina. Cuando los algoritmos

son parte de la respuesta, suscitan una serie de objeciones; abordamos aquí algunas de ellas. Aun así, el nivel actual de ruido es inaceptable. No obstante, instamos a que tanto las organizaciones privadas como las públicas lleven a cabo auditorías del ruido y se esfuercen, con una seriedad sin precedentes, por reducir el ruido. Si lo hicieran, las organizaciones podrían reducir la arbitrariedad generalizada (y reducir costes en muchos ámbitos).

Con esta aspiración concluimos cada capítulo haciendo unas breves propuestas con forma de citas ficticias. Pueden utilizarse estas propuestas tal como figuran o adaptarlas a cualquier asunto que interese, esté relacionado con la salud, la seguridad, la educación, el dinero, el empleo, el ocio o cualquier otro tema. Entender el problema del ruido y tratar de resolverlo es una tarea constante y una empresa colectiva. Todos tenemos oportunidades de contribuir a esa labor. Este libro se ha escrito con la esperanza de que podamos aprovechar esas oportunidades.

PRIMERA PARTE

Buscando el ruido

No es aceptable que individuos similares, condenados por el mismo delito, terminen con sentencias notablemente diferentes; por ejemplo, cinco años de cárcel para uno y libertad condicional para otro. Y, sin embargo, en muchos lugares ocurren estas cosas. Sin duda, el sistema de justicia penal también está impregnado de sesgos, pero en el capítulo 1 nos vamos a centrar en el ruido, y, en particular, en lo que ocurrió cuando un conocido juez llamó la atención sobre él, lo consideró escandaloso y lanzó una cruzada que, en cierto sentido, cambió el mundo (aunque no lo suficiente). Nuestra historia se sitúa en Estados Unidos, pero estamos seguros de que se pueden contar (y se contarán) historias parecidas de muchas otras naciones. En algunas de ellas es probable que el problema del ruido sea aún mayor que en Estados Unidos. Utilizamos el ejemplo de las sentencias para demostrar en parte que el ruido puede ser causa de grandes injusticias.

Las sentencias penales son especialmente dramáticas, pero del mismo modo nos preocupa el sector privado, donde también hay mucho en juego. Para ilustrar este punto, en el capítulo 2 nos detendremos en una gran compañía de seguros. Allí, los agentes de seguros cumplen la tarea de fijar las primas de los seguros para potenciales clientes, y los tasadores deben juzgar el valor de las reclamaciones. Se podría predecir que estas tareas serían sencillas y mecánicas, y que diferentes profesionales establecerían casi las mismas cuantías. Hemos realizado un experimento cuidadosamente diseñado —una auditoría del ruido— para comprobar esa predicción. Los resultados nos sorprendieron, pero sobre todo asombraron y consternaron a la dirección de la compañía. Nos dimos cuenta de que un gran volumen de ruido les estaba costando mucho dinero. Utilizamos este ejemplo para demostrar que el ruido puede ocasionar grandes pérdidas económicas.

Ambos ejemplos suponen estudios de un gran número de personas que realizan un gran número de juicios. Sin embargo, muchos juicios importantes son más «singulares» que repetitivos: cómo tratar una oportunidad de negocio en apariencia única; dudar sobre si lanzar un nuevo producto; cómo hacer frente a una pandemia; dudar sobre si contratar a alguien que no da el perfil estándar. ¿Se puede encontrar ruido en

decisiones sobre situaciones únicas como estas? Es tentador pensar que allí está ausente. Al fin y al cabo, el ruido es una variabilidad no deseada, y ¿cómo puede haber variabilidad en decisiones singulares? En el capítulo 3 tratamos de responder a esta pregunta. El juicio que se hace incluso en una situación aparentemente única es uno en una nube de posibilidades. Allí también encontraremos gran cantidad de ruido.

El tema que aflora en estos tres capítulos puede resumirse en una frase, que será un tema clave del presente libro: «Dondequiera que haya juicio, hay ruido (y más del que se piensa)». Averigüemos cuánto.

1
Crimen y castigo ruidoso

Supongamos que alguien ha sido condenado por cometer un delito —robo en comercios, posesión de heroína, asalto o robo a mano armada—. ¿Cuál es la sentencia más probable?

La respuesta no debería depender del juez particular al que se asigne el caso, de si fuera hace calor o frío o de si un equipo deportivo local ganó el día anterior. Sería escandaloso que tres personas similares, condenadas por el mismo delito, recibieran penas radicalmente diferentes: libertad condicional para una, dos años de cárcel para otra y diez años de cárcel para la restante. Y, sin embargo, esta clase de arbitrariedades puede encontrarse en muchas naciones (y no solo en un pasado lejano, sino también hoy).

En todo el mundo, los jueces han mostrado durante mucho tiempo una más que notable discrecionalidad para decidir las sentencias apropiadas. En muchas naciones, los expertos han celebrado esa discrecionalidad y la han considerado justa y humana. Insistían en que las sentencias penales debían basarse en una serie de factores que no solo especificaban el delito, sino también el carácter y las circunstancias del acusado. La sentencia a medida estaba a la orden del día. Si los jueces se vieran obligados por unas normas, los delincuentes serían tratados de una manera deshumanizada; no serían considerados como individuos únicos con derecho a que se les tengan en cuenta los detalles de su situación. A muchos les parecía que la propia idea del juicio justo demandaba un discrecionalidad judicial sin trabas.

En la década de 1970, el entusiasmo universal por la discrecionalidad judicial comenzó a derrumbarse por una simple razón: la alarmante evidencia de ruido. En 1973, un célebre juez, Marvin Frankel, llamó la atención del público sobre el problema. Antes de ser juez, Frankel fue un defensor de la libertad de expresión y un apasionado valedor de los derechos humanos que colaboró en la fundación del Comité de Abogados por los Derechos Humanos (una organización conocida hoy con el nombre de Human Rights First).

Frankel podía ser implacable, y estaba indignado con el ruido en el sistema de justicia penal. Así es como describía su motivación:

> Si un acusado de atraco a un banco federal era condenado, podía ser sentenciado a un máximo de veinticinco años. Esto significa cualquier condena de cero a veinticinco años. Y, cuando se fijó el número, pronto me percaté de que dependía menos del caso o del individuo acusado que del juez; es decir, de las opiniones, predilecciones e inclinaciones del juez. Y así, el mismo acusado en el mismo caso podía recibir sentencias muy diferentes dependiendo de qué juez se encargaba del caso.[1]

Frankel no ofreció ningún tipo de análisis estadístico para respaldar su argumentación, pero sí una serie de convincentes anécdotas que mostraban disparidades injustificadas en el tratamiento de personas muy similares. Dos hombres, ninguno de los cuales tenía antecedentes penales, fueron condenados por haber falsificado cheques por valor de 58,40 y 35,20 dólares, respectivamente. El primero, a quince *años* de prisión, y el segundo, a treinta *días* de prisión. Por actos de malversación de fondos similares, un hombre fue condenado a ciento diecisiete *días* de prisión, mientras que otro lo fue a veinte *años*. Citando numerosos casos de este tipo, Frankel deploraba los que llamó «poderes casi totalmente incontrolados y arrolladores»[2] de jueces federales con el resultado de «crueldades arbitrarias perpetradas diariamente»,[3] que consideraba inaceptables en un «gobierno de leyes, no de hombres».[4]

Frankel apeló al Congreso para que pusiera fin a aquella «discriminación», palabra con la que definió esas crueldades arbitrarias. Con ese término se refería sobre todo al ruido en forma de inexplicables variaciones en las sentencias. Sin embargo, también le preocupaban los sesgos en forma de disparidades por motivos raciales y socioeconómicos. Para combatir tanto el ruido como el sesgo, instó a que no se permitieran diferencias de trato a los acusados de delitos, a menos que pudieran «justificarse mediante pruebas pertinentes capaces de formularse y aplicarse con suficiente objetividad para garantizar que los resultados sean más determinantes que los idiosincrásicos ucases de determinados funcionarios, jueces u otros».[5] (La expresión «idiosincrásicos ucases» es un tanto esotérica; Frankel se refería a los que parecían decretos personales.) Y mucho más que eso: Frankel abogó por una reducción del ruido mediante un «perfil detallado o lista de control de factores que incluiría, siempre que

fuera posible, alguna forma de gradación numérica u otra que fuese objetiva».[6]

Escribiendo a principios de la década de 1970, no llegó a defender lo que llamó «desplazamiento de las personas por las máquinas», pero, sorprendentemente, estuvo bastante cerca de hacerlo. Creía que «el imperio de la ley demanda un cuerpo de reglas impersonales, aplicables a todos los niveles, que sean vinculantes para los jueces y para otros responsables». Defendió explícitamente el uso de «computadoras como refuerzo de un pensamiento disciplinado en las sentencias».[7] También recomendó la creación de una «comisión de sentencias».[8]

El libro de Frankel fue uno de los más influyentes de toda la historia del derecho penal no solo en Estados Unidos, sino también en todo el mundo. En su trabajo había cierto grado de informalidad. Fue arrollador, pero era una visión impresionista. Para probar la realidad del ruido, unos cuantos investigadores se dedicaron de inmediato a examinar el nivel de ruido de las sentencias penales.

En 1974 se llevó a cabo un primer estudio a gran escala de este tipo, presidido por el propio juez Frankel. Se pidió a cincuenta jueces de diferentes distritos que dictaran sentencias para acusados en casos hipotéticos resumidos en informes idénticos previos a las sentencias. La conclusión fue, básicamente, que «la ausencia de consenso era la norma»[9] y que las variaciones entre las penas eran «asombrosas».[10] Un vendedor de heroína podía ser encarcelado de uno a diez años, dependiendo del juez.[11] Las penas para un atracador de bancos oscilaban entre cinco y dieciocho años de prisión.[12] El estudio encontró que, en un caso de extorsión, las sentencias variaban desde la friolera de veinte años de prisión y una multa de sesenta y cinco mil dólares hasta solo tres años de prisión y ninguna multa.[13] Y lo más sorprendente de todo: en dieciséis de los veinte casos, no hubo unanimidad sobre si era apropiado un encarcelamiento.

A este estudio le siguió una serie de otros más, todos los cuales encontraron similares e inaceptables niveles de ruido. En 1977, por ejemplo, William Austin y Thomas A. Williams III llevaron a cabo una encuesta entre cuarenta y siete jueces, a los que pidieron que respondieran a los mismos cinco casos, todos ellos de delitos de nivel relativamente bajo.[14] Todas las descripciones de los casos incluían resúmenes de la

información utilizada por los jueces en sentencias reales, incluidos la acusación, el testimonio, los antecedentes penales previos (si los había), la extracción social y pruebas relacionadas con el carácter. El hallazgo clave fue una «disparidad sustancial». En un caso de robo, por ejemplo, las sentencias recomendadas oscilaban entre cinco años y solo treinta días de prisión (más una multa de cien dólares). En un caso de posesión de marihuana, algunos jueces recomendaron penas de prisión; otros, la libertad condicional.

En un estudio mucho más amplio, realizado en 1981, participaron doscientos ocho jueces federales, a los que se expusieron los mismos dieciséis casos hipotéticos.[15] Sus principales hallazgos fueron pasmosos:

> En solo tres de los dieciséis casos hubo un acuerdo unánime en imponer una pena de prisión. Sin embargo, cuando la mayoría de los jueces estuvieron de acuerdo en que una pena de prisión era apropiada, hubo una variación sustancial en la duración de las penas de prisión recomendadas. En un caso de fraude en el que la pena media de prisión era de 8,5 años, la pena más larga fue de cadena perpetua. En otro caso, la pena media de prisión fue de 1,1 años, pero la pena de prisión más larga recomendada fue de quince años.

Por muy reveladores que sean estos estudios, que se valen de experimentos muy controlados, es casi seguro que subestiman la magnitud del ruido en el mundo real de la justicia penal. Los jueces de la vida real tienen delante mucha más información de la que los participantes en el estudio tenían en las historias cuidadosamente especificadas de esos experimentos. Parte de esta información adicional es, por supuesto, pertinente, pero también hay abundantes pruebas de que información irrelevante, en forma de factores nimios y aparentemente aleatorios, puede dar lugar a grandes diferencias en los resultados. Por ejemplo, se ha comprobado que es más probable que los jueces concedan la libertad condicional al comienzo del día o después de una pausa para comer que inmediatamente antes de dicha pausa. Si los jueces están hambrientos, son más duros.[16]

Un estudio realizado sobre miles de decisiones tomadas por tribunales de menores encontró que, cuando el equipo de fútbol americano local pierde un partido el fin de semana, los jueces toman decisiones más duras el lunes (y, en menor medida, el resto de la semana).[17] Los acusados negros se llevaron, con diferencia, la peor parte. Un estudio diferente examinó

durante tres décadas un millón y medio de decisiones judiciales y encontró que, en los días que seguían a una derrota del equipo de fútbol americano de la ciudad, los jueces eran más severos que durante los días siguientes a una victoria.[18]

Un estudio de seis millones de decisiones tomadas por jueces en Francia durante doce años encontró que estos eran más indulgentes con los acusados en el día de su cumpleaños.[19] (En el cumpleaños del acusado; sospechamos que los jueces podrían ser más indulgentes también en sus propios cumpleaños, aunque, por lo que sabemos, esa hipótesis no ha sido contrastada.) Incluso algo tan irrelevante como la temperatura exterior puede influir en los jueces.[20] Una revisión de doscientas siete mil decisiones que tomaron los tribunales de inmigración en cuatro años encontró un efecto significativo de las variaciones diarias de temperatura: cuando fuera hacía calor, era menos probable obtener asilo. Si el solicitante sufre persecución política en su país de origen y busca asilo en otro, debe mantener la esperanza y tal vez rezar por que haga un día más fresco.

REDUCIENDO EL RUIDO EN LAS SENTENCIAS

En la década de 1970, los argumentos de Frankel y los hallazgos empíricos que los respaldaban llamaron la atención de Edward M. Kennedy, hermano del presidente asesinado John F. Kennedy y uno de los miembros más influyentes del Senado de Estados Unidos. Kennedy estaba consternado y horrorizado. Ya en 1975, propuso una ley de reforma de las sentencias; no fue a ninguna parte. Sin embargo, Kennedy fue implacable. Señalando las pruebas, continuó presionando año tras año para que se promulgara esa ley. En 1984 lo consiguió. En respuesta a las pruebas de variabilidad injustificada, el Congreso promulgó su Ley de Reforma de las Sentencias de 1984.

La nueva ley tenía por objeto disminuir el ruido en el sistema reduciendo «la discrecionalidad irrestricta que la ley confiere a jueces y autoridades decisorias de la libertad condicional como responsables de imponer e implementar las sentencias».[21] Los miembros del Congreso hicieron una particular referencia a la disparidad de sentencias, que consideraban «injustificadamente amplia», citando en concreto los hallazgos que probaban que, en la zona de Nueva York, las penas para casos reales

idénticos podían oscilar entre tres y veinte años de prisión.[22] La ley creó la Comisión de Sentencias de Estados Unidos, tal y como había recomendado el juez Frankel, cuya principal tarea era clara: marcar a las sentencias directrices que serían obligatorias y que establecerían un rango restringido para las sentencias penales.

Al año siguiente, la comisión estableció esas directrices, basadas en general en las sentencias medias por delitos similares en un análisis de diez mil casos reales. El juez del Tribunal Supremo Stephen Breyer, que participó intensamente en el proceso, defendió la práctica del pasado señalando el desacuerdo irresoluble en el seno de la comisión: «¿Por qué la comisión no se reunió para razonar este asunto y no solo para hacer historia? La respuesta es breve: no pudimos. No pudimos porque hay buenos argumentos por todas partes apuntando en direcciones opuestas [...] Intenten enumerar todos los delitos que existen por orden de punibilidad [...] Reúnan luego los resultados de sus amigos y observen si todos coinciden. Les diré que no lo hacen».[23]

Según las directrices, los jueces debían considerar dos factores para establecer las sentencias: el delito y el historial delictivo del acusado. A los delitos se les asignaban cuarenta y tres «niveles delictivos», dependiendo de su gravedad. En el historial delictivo del acusado contaba sobre todo el número y la gravedad de sus condenas anteriores. Una vez reunidos el delito y los antecedentes, las directrices ofrecían un rango relativamente reducido de sentencias, con la autorización de que la parte superior de la horquilla excediese a la inferior en un máximo de seis meses o un 25 por ciento. Se permitía a los jueces salirse por completo de esa horquilla en consideración a las circunstancias agravantes o atenuantes, pero era algo que debían justificar ante un tribunal de apelación.[24]

Aunque las directrices son obligatorias, no son del todo rígidas. No van tan lejos como el juez Frankel quería. Ofrecen a los jueces un espacio importante para maniobrar. No obstante, distintos estudios que utilizan una variedad de métodos y se centran en diferentes periodos, han llegado a la misma conclusión: las directrices disminuyen el ruido. En términos más técnicos, «redujeron la variación neta de las sentencias atribuibles a la casualidad de la identidad del juez que dictó sentencia».[25]

El estudio más elaborado provino de la propia comisión.[26] Esta comparó las sentencias en los casos de atraco a bancos, tráfico de cocaína,

tráfico de heroína y malversación de fondos en 1985 (antes de que las directrices entraran en vigor) con las dictadas entre el 19 de enero de 1989 y el 30 de septiembre de 1990. Se emparejó a los delincuentes con respecto a los factores que se consideraban pertinentes para dictar sentencia con arreglo a las directrices. Por cada delito, las variaciones entre los jueces fueron mucho menores en el último periodo, después de aprobarse la Ley de Reforma de las Sentencias.

Según otro estudio, la diferencia esperada en la duración de las condenas entre los jueces fue del 17 por ciento, o de 4,9 meses, en 1986 y 1987. Ese número descendió al 11 por ciento, o a 3,9 meses, entre 1988 y 1993.[27] Un estudio independiente, que abarcó diferentes periodos, encontró un éxito similar en la reducción de las disparidades entre los jueces, definidas como las diferencias en las sentencias medias entre jueces que tenían una cantidad similar de casos.[28]

A pesar de estos hallazgos, las directrices se encontraron con una tormenta de críticas. Algunas personas, entre ellas muchos jueces, pensaban que algunas sentencias eran demasiado severas (fruto del sesgo, no del ruido). Para nuestros propósitos, una objeción mucho más interesante, proveniente de numerosos jueces, era que las directrices eran muy injustas, porque prohibían a los jueces tener en cuenta de forma adecuada las particularidades del caso. El precio de reducir el ruido era tomar decisiones inaceptablemente mecánicas. La profesora de Derecho de Yale Kate Stith y el juez federal José Cabranes escribieron que «la necesidad no es de ceguera, sino de conocimiento, de equidad», lo cual «solo puede darse en un juicio que tenga en cuenta las complejidades del caso particular».[29]

Esta objeción ocasionó enérgicas impugnaciones a las directrices, unas basadas en la ley y otras fundadas en la política adoptada. Esas impugnaciones se obviaron hasta que, por razones del todo ajenas al debate que aquí resumimos, el Tribunal Supremo anuló las directrices en 2005.[30] Como resultado del fallo de este tribunal, las directrices se quedaron en algo meramente consultivo. Es revelador que los jueces federales se quedaran mucho más contentos después de que el Tribunal Supremo tomara esa decisión. El 75 por ciento prefería el régimen consultivo, mientras que solo el 3 por ciento pensaba que la obligatoriedad era mejor.[31]

¿Cuáles habían sido los efectos de cambiar las directrices de obligatorias a consultivas? Crystal Yang, profesora de Derecho en Harvard, investigó

esta cuestión no con un experimento o con una encuesta, sino con un conjunto masivo de datos de sentencias reales dictadas contra casi cuatrocientos mil acusados. Su principal hallazgo fue que, en múltiples mediciones, las disparidades entre los jueces aumentaron de forma significativa después de 2005. Cuando las directrices eran obligatorias, los acusados que habían sido condenados por un juez relativamente severo fueron sentenciados con 2,8 meses más que si hubieran sido condenados por un juez del promedio. Cuando las directrices fueron meramente consultivas, la disparidad se duplicó. En franca consonancia con lo señalado por el juez Frankel hacía cuarenta años, Yang escribió que sus «hallazgos plantean grandes cuestiones de equidad, ya que la identidad del juez sentenciador asignado contribuye de forma significativa al tratamiento dispar de delincuentes similares condenados por delitos similares».[32]

Después de que las directrices fueran consultivas, los jueces se volvieron más propensos a basar sus sentencias en sus valores personales. Las directrices obligatorias reducen tanto el sesgo como el ruido. Tras la decisión del Tribunal Supremo, se produjo un aumento significativo de la disparidad entre las sentencias de los acusados afroamericanos y las de los blancos condenados por los mismos delitos. A su vez, las juezas se volvieron más propensas a una mayor discrecionalidad a favor de la indulgencia. Al igual que ocurría con los jueces nombrados por presidentes demócratas.

Tres años después de fallecer el juez Frankel en 2002, la abolición de las directrices produjo un retorno a algo más parecido a la que fue su pesadilla: la ley sin orden.

La historia de la batalla del juez Frankel por implantar unas directrices para las sentencias permite vislumbrar varios de los puntos clave que trataremos en este libro. El primero de ellos es que el acto de juzgar es difícil, porque el mundo es un lugar complicado e incierto. Esta complejidad es obvia en la judicatura, y también existe en la mayoría de las situaciones que requieren un juicio profesional. En líneas generales, estas situaciones incluyen juicios emitidos por médicos, enfermeras, abogados, ingenieros, profesores, arquitectos, ejecutivos de Hollywood, miembros de comités de contratación, editores de libros, ejecutivos de toda clase de empresas y

directivos de equipos deportivos. El desacuerdo es inevitable cuando se emite un juicio.

El segundo punto es que el alcance de estos desacuerdos es mucho mayor de lo que esperamos. Aunque pocas personas se oponen al principio de discrecionalidad judicial, casi todas desaprueban la magnitud de las disparidades que produce. El «ruido del sistema», es decir, la variabilidad no deseada en juicios que idealmente deberían ser idénticos, puede generar una injusticia rampante, elevados costes económicos y muchas clases de errores.

El tercer punto es que el ruido puede reducirse. El método por el que abogaba Frankel, implementado por la Comisión de Sentencias —reglas y directrices—, es uno de los varios posibles que logran reducir el ruido. Otros métodos se adaptan mejor a otros tipos de juicio. Algunos de los métodos adoptados para reducir el ruido pueden reducir simultáneamente el sesgo.

El cuarto punto es que los intentos de reducción del ruido suscitan a menudo objeciones y se encuentran con serias dificultades. Esto debe también abordarse, o la lucha contra el ruido fracasará.

SOBRE EL RUIDO EN LAS SENTENCIAS

- «Los experimentos muestran grandes disparidades entre los jueces en las sentencias que recomendaban para casos idénticos. Esta variabilidad no puede ser justa. La sentencia para un acusado no debe depender del juez al que se le asigne el caso.»
- «Las sentencias penales no han de obedecer al estado de ánimo del juez durante la vista ni a la temperatura exterior.»
- «Las directrices son una forma de abordar esta cuestión. Sin embargo, a mucha gente no le gustan, porque limitan la discrecionalidad judicial, que podría ser necesaria para garantizar la equidad y la exactitud. Después de todo, cada caso es único, ¿no?»

2
Un sistema ruidoso

Nuestro encuentro inicial con el ruido, así como lo que primero despertó nuestro interés por este asunto, no fue tan dramático como un roce con el sistema de justicia penal. En realidad, el encuentro fue una especie de accidente en el que estuvo envuelta una compañía de seguros que había contratado a la consultoría con la que dos de nosotros estábamos asociados.

El tema de los seguros no es ciertamente del agrado de todos. Sin embargo, lo que encontramos nos muestra la magnitud del problema del ruido en una organización con fines de lucro que puede perder mucho con las decisiones ruidosas. Nuestra experiencia con la compañía de seguros ayuda a explicar por qué ese problema pasa tantas veces inadvertido y nos enseña lo que se puede hacer con él.

Los ejecutivos de la compañía de seguros sopesaban el valor potencial de un esfuerzo por aumentar la consistencia —por reducir el ruido— en los juicios de las personas que tomaban importantes decisiones financieras en nombre de la empresa. Todas esas personas estaban de acuerdo en que la consistencia era deseable. Todas estuvieron también de acuerdo en que esos juicios nunca podrían ser totalmente consistentes, porque son informales y, en parte, subjetivos. Algo de ruido era inevitable.

El desacuerdo surgió cuando se consideró su magnitud. Los ejecutivos no creían que el ruido pudiera ser un problema sustancial para su compañía. Sin embargo, se dignaron de mutuo acuerdo a resolver la cuestión mediante una especie de experimento bastante simple que llamaremos una «auditoría del ruido». El resultado fue una sorpresa para ellos. También resultó ser una ilustración ideal del problema del ruido.

UNA LOTERÍA QUE GENERA RUIDO

Muchos profesionales de cualquier gran empresa están autorizados a emitir juicios que vinculan a la empresa. Por ejemplo, la referida compañía de seguros emplea a un número considerable de agentes de seguros que

ofrecen primas para riesgos financieros, como asegurar a un banco contra pérdidas por fraude o comercio deshonesto. También emplea a muchos tasadores de reclamaciones que pronostican el coste de futuras reclamaciones y las negocian con los reclamantes si se producen disputas.

En cualquier gran rama de la compañía hay varios agentes cualificados. Cuando se solicita una cotización, se puede asignar a cualquiera que esté disponible para prepararla. El agente particular que determinará una cotización es seleccionado por sorteo.

El valor exacto de la cuota tiene consecuencias importantes para la compañía. Una prima alta es ventajosa si se acepta la cuota, pero se corre el riesgo de la pérdida de ingresos frente a un competidor. Es más probable que se acepte una prima baja, aunque es menos ventajosa para la compañía. Para cualquier riesgo se fija un precio de Ricitos de Oro que es justo —ni muy alto ni muy bajo—, y lo más probable es que el juicio promedio de un gran grupo de profesionales no esté muy lejos de él. Los precios superiores o inferiores a esta cuantía son costosos; así es como la variabilidad de los juicios ruidosos perjudica el balance final.

El trabajo de los tasadores también afecta a las finanzas de la compañía. Por ejemplo, supongamos que se presenta una reclamación en nombre de un trabajador (el reclamante) que ha perdido definitivamente el uso de su mano derecha en un accidente laboral. Se asigna un tasador a la reclamación, como en el caso del agente de seguros, porque resulta que está disponible. El tasador reúne los hechos del caso y hace una estimación de su coste final para la compañía. El mismo tasador se encarga entonces de negociar con el representante del reclamante para garantizarle que este recibirá los beneficios prometidos en la póliza, protegiendo al mismo tiempo a la compañía de desembolsos excesivos.

La primera estimación es importante, porque establece un objetivo implícito para el tasador en futuras negociaciones con un reclamante. La compañía de seguros también está legalmente obligada a reservar el coste previsto de cada reclamación (es decir, a tener dinero suficiente para poder pagarla). Aquí también hay un valor de Ricitos de Oro desde la perspectiva de la empresa. El acuerdo no está garantizado, ya que, por la otra parte, hay un abogado para el reclamante que puede optar por acudir a los tribunales si lo que se le ofrece a este es mísero. Por otro lado, una reserva en exceso generosa puede dar al tasador demasiada libertad para aceptar

reclamaciones frívolas. El juicio del tasador es importante para la compañía, y aún más para el reclamante.

Hemos elegido la palabra «lotería» para subrayar el papel del azar en la selección de un agente o de un tasador. En el funcionamiento normal de la compañía, se asigna un solo profesional a un caso, y nadie puede saber lo que habría ocurrido si se hubiera seleccionado a otro colega suyo.

Las loterías tienen su sitio, y no son necesariamente injustas. Hay sorteos aceptables utilizados para repartir «bienes», como los cursos en algunas universidades, o «males», como el servicio militar. Sirven para algo. Sin embargo, las loterías de los juicios, de las que aquí hablamos, no reparten nada. Solo producen incertidumbre. Imagínese una compañía de seguros cuyos agentes no generan ruido y fijan la prima óptima, pero luego interviene un mecanismo de azar para modificar la cuota que el cliente tiene a la vista. Evidentemente, no habría ninguna justificación para semejante lotería. Tampoco se justifica un sistema en el que el resultado dependa de la identidad de la persona elegida al azar para emitir un juicio profesional.

LAS AUDITORÍAS DEL RUIDO REVELAN EL RUIDO DEL SISTEMA

La lotería que elige a un juez concreto para dictar una sentencia penal, o a un único tirador para representar a un equipo, crea variabilidad, pero esta variabilidad no se ve. Una auditoría del ruido —del tipo que se realizó a los jueces federales en relación con las sentencias— es una forma de revelar el ruido. En esta auditoría, el mismo caso es evaluado por muchos individuos, y la variabilidad de sus respuestas se hace visible.

Los juicios de agentes de seguros y tasadores de reclamaciones se prestan especialmente bien a este ejercicio, porque sus decisiones se basan en información escrita. Para preparar la auditoría del ruido, los ejecutivos de la compañía hicieron descripciones detalladas de cinco casos representativos por cada grupo (agentes y tasadores). Se pidió a estos empleados que evaluaran dos o tres casos cada uno trabajando de forma independiente. No se les dijo que el propósito del estudio era examinar la variabilidad de sus juicios.[1]

Antes de seguir leyendo, tal vez el lector desee pensar su propia respuesta a la siguiente pregunta: en una compañía de seguros bien gestionada, si seleccionaran al azar dos agentes o tasadores de reclamaciones cualificados,

¿cuán diferentes esperaría que fueran sus estimaciones para el mismo caso? Más en concreto, ¿cuál sería la diferencia entre las dos estimaciones expresada en un porcentaje de su promedio?

Pedimos sus respuestas a numerosos ejecutivos de la compañía, y en los años siguientes hemos obtenido estimaciones de una gran variedad de personas de diferentes profesiones. Sorprendentemente, una respuesta era mucho más popular que todas las demás. La mayoría de los ejecutivos de la compañía de seguros estimaron el 10 por ciento o menos. Cuando preguntamos a ochocientos veintiocho directores ejecutivos (CEO) y a ejecutivos de alto nivel de diversas industrias cuánta variación esperaban encontrar en juicios similares, el 10 por ciento fue también la respuesta media y la más frecuente (la segunda más popular fue el 15 por ciento). Una diferencia del 10 por ciento significaría, por ejemplo, que uno de los dos agentes fijó una prima de nueve mil quinientos dólares, mientras que la del otro fue de diez mil quinientos dólares. No es una diferencia insignificante, pero se puede esperar que una organización la tolere.

Nuestra auditoría del ruido encontró diferencias mucho mayores. Según nuestra medición, la diferencia media en las primas fue del 55 por ciento, unas cinco veces mayor de lo que la mayoría de las personas, incluidos los ejecutivos de la compañía, esperaban. Este resultado significa que, por ejemplo, cuando un agente fija una prima en nueve mil quinientos dólares, el otro no la fija en diez mil quinientos dólares, sino en dieciséis mil setecientos dólares. Para los tasadores de reclamaciones, la ratio media fue del 43 por ciento. Subrayamos que estos resultados son medianas: en la mitad de los pares de casos, la diferencia entre los dos juicios fue aún mayor.

Los ejecutivos a los que informamos de los resultados de la auditoría del ruido no tardaron en darse cuenta de que el simple volumen de ruido representaba un problema costoso. Un alto ejecutivo estimó que, para la compañía, el coste anual del ruido en la suscripción de seguros —contando las pérdidas en el negocio debidas a las cuotas excesivas y las pérdidas provocadas por los contratos de bajo precio— era de cientos de millones de dólares.

Nadie pudo decir con precisión cuánto había de error (o de sesgo), porque nadie podía saber con seguridad el valor de Ricitos de Oro para cada caso. Sin embargo, no había necesidad de ver la diana para medir la dispersión en

el dorso del blanco y darse cuenta de que la variabilidad era un problema. Los datos mostraban que el precio que se pide que pague un cliente dependía en un grado turbador de la lotería del empleado elegido para ocuparse de esa transacción, y a los clientes no les gustaría oír que los inscribieron en esa lotería sin su consentimiento. En términos más generales, las personas que tratan con organizaciones confían en un sistema cuyos juicios son consistentes. No esperan ese ruido del sistema.

VARIABILIDAD INDESEADA *VERSUS* DIVERSIDAD DESEADA

Una característica definitoria de ruido del sistema es que es «indeseado», y debemos subrayar aquí que la variabilidad en los juicios no siempre es indeseada.

Consideremos las cuestiones de preferencia o de gusto. Si diez críticos de cine ven la misma película, si diez catadores califican el mismo vino o si diez personas leen la misma novela, no esperamos que tengan la misma opinión. La variedad de gustos es bienvenida y totalmente esperada. Nadie querría vivir en un mundo en el que todas las personas tuvieran los mismos gustos y las mismas antipatías. (Bueno, casi nadie.) Sin embargo, la variedad de gustos puede ayudar a explicar errores cuando se confunde un gusto personal con un juicio profesional. Si un productor de cine decide seguir adelante con un proyecto inusual (sobre, digamos, el auge y la desaparición del teléfono de disco) porque a él personalmente le gusta el guion, podría cometer un gran error si a nadie más le gusta.

También se espera y es bien recibida la variabilidad de los juicios en una situación de competencia en la que los mejores juicios serán recompensados. Cuando varias empresas (o varios equipos dentro de la misma organización) compiten para generar soluciones innovadoras al mismo problema del cliente, no se desea que se centren en el mismo enfoque. Lo propio ocurre cuando múltiples equipos de investigadores abordan un problema científico, como el desarrollo de una vacuna: queremos que lo miren desde diferentes ángulos. Hasta los analistas se comportan a veces como jugadores que compiten. El analista que llama por su nombre a una recesión que nadie más ha previsto seguramente ganará fama, mientras que el que nunca se aleja del consenso permanecerá en la

oscuridad. En tales situaciones, la variabilidad de ideas y juicios es también bienvenida, porque la variación es solo el primer paso. En una segunda fase, los resultados de estos juicios se enfrentarán entre sí y triunfarán los mejores. En un mercado, como en la naturaleza, la selección no puede funcionar sin variación.

Las cuestiones de gusto y los escenarios de competición plantean interesantes problemas de juicio, pero nos centraremos en los juicios en los que la variabilidad es indeseada. El ruido de los sistemas es un problema de los sistemas, que son organizaciones, no mercados. Cuando los operadores de bolsa hacen diferentes estimaciones del valor de una acción, unos ganarán dinero y otros, no. Los desacuerdos crean mercados. Sin embargo, si uno de esos operadores fue elegido al azar para hacer esa evaluación en nombre de su empresa y encontramos que sus colegas de la misma empresa han hecho evaluaciones muy diferentes, entonces hay ruido del sistema en la empresa, y eso es un problema.

Un elegante ejemplo de este problema se nos mostró cuando presentamos nuestros hallazgos a los directivos de una empresa de gestión de activos, lo cual les dio pie para llevar a cabo su propia auditoría exploratoria del ruido. Pidieron a cuarenta y dos inversores experimentados de la empresa que estimaran el valor justo de una acción (el precio al que a los inversores les resultaría indiferente comprar o vender). Los inversores basaron su análisis en una descripción del negocio de solo una página; los datos incluían ganancias y pérdidas simplificadas, el balance general, resúmenes del flujo de caja en los tres últimos años y proyecciones para los dos próximos. El ruido mediano, medido de la misma manera que en la compañía de seguros, fue del 41 por ciento. Diferencias tan grandes entre los inversores de una misma empresa, que utilizan los mismos métodos de evaluación, no pueden ser una buena noticia.

Siempre que la persona que emite un juicio es seleccionada al azar de un grupo de individuos igualmente cualificados, como es el caso en esta empresa de gestión de activos, en el sistema de justicia penal y en la compañía de seguros arriba mencionada, el ruido es un problema. El ruido del sistema asedia a muchas organizaciones similares: un proceso de asignación efectivamente aleatorio suele decidir qué médico atenderá a un paciente en un hospital, qué juez verá un caso en un tribunal, qué examinador de patentes estudiará una solicitud, qué representante del

servicio de atención al cliente escuchará una queja, etcétera. La variabilidad no deseada en estos juicios puede causar serios problemas, entre ellos la pérdida de dinero y una injusticia rampante.

Una equivocación frecuente sobre la variabilidad no deseada en los juicios es pensar que esta no importa, porque los errores aleatorios supuestamente se cancelan entre sí. Sin duda, los errores positivos y negativos en un juicio sobre el mismo caso tenderán a anularse mutuamente, y discutiremos en detalle cómo puede utilizarse esta propiedad para reducir el ruido. Pero los sistemas ruidosos no producen múltiples juicios sobre el mismo caso; producen juicios ruidosos sobre diferentes casos. Si el precio de una póliza de seguros es demasiado alto y el de otra demasiado bajo, el precio podrá parecer correcto de promedio, pero la compañía de seguros habrá cometido dos costosos errores. Si dos malhechores que deberían ser condenados a cinco años de prisión reciben sentencias de tres y siete años, nadie podrá decir que se ha hecho justicia. En los sistemas ruidosos, los errores no se cancelan: se suman.

LA ILUSIÓN DE ACUERDO

Una nutrida literatura de varias décadas ha documentado el ruido en el juicio profesional. Como estábamos al tanto de ella, los resultados de la auditoría del ruido hecha a la compañía de seguros no nos sorprendieron. Lo que nos sorprendió fue la reacción de los ejecutivos a los que informamos de nuestros hallazgos: nadie en la compañía esperaba nada parecido a la cantidad de ruido que habíamos observado. Nadie cuestionó la validez de la auditoría y nadie afirmó que la cantidad de ruido que habíamos observado era aceptable. Sin embargo, parecía que el problema del ruido —y sus grandes costes— era algo nuevo para la organización. El ruido era como una fuga en el sótano. Se toleraba no porque se la considerase aceptable, sino porque había pasado inadvertida.

¿Cómo es esto posible? ¿Cómo pueden unos profesionales que desempeñan la misma función y están en la misma oficina diferir tanto unos de otros sin advertirlo? ¿Cómo podían los ejecutivos no haber observado este hecho, que entendieron que era una importante amenaza a la labor y la reputación de su compañía? Llegamos a la conclusión de que el problema del ruido de un sistema a menudo no se reconoce en las organizaciones, y

que la extendida inatención al ruido es tan interesante como su frecuencia. La auditoría del ruido indicaba que profesionales respetados —y las organizaciones que los empleaban— mantenían una «ilusión de acuerdo», cuando en realidad existía desacuerdo entre sus juicios profesionales diarios.

Para empezar a entender cómo se produce la ilusión de acuerdo, pongámonos en el lugar de un agente de seguros en un día de trabajo normal. Tiene más de cinco años de experiencia, sabe que está bien considerado entre sus colegas, a los que respeta y valora. Sabe que es bueno en su trabajo. Tras analizar a fondo los complejos riesgos que enfrenta una empresa financiera, su conclusión es que una prima de doscientos mil dólares es apropiada. El problema es complicado, pero no muy diferente de los que resuelve todos los días de la semana.

Imaginemos ahora que los colegas de su oficina han recibido la misma información y han evaluado un riesgo de la misma manera. ¿Podría creer que la mitad de ellos, o más, establecen una prima que es superior a doscientos cincuenta y cinco mil dólares o inferior a ciento cuarenta y cinco mil dólares? Esto es difícil de aceptar. Cabe sospechar que los agentes de seguros que oyeron hablar de la auditoría del ruido y aceptaron su validez nunca creyeron en verdad que sus conclusiones se aplicaran a ellos personalmente.

La mayoría de los humanos vivimos la mayor parte del tiempo con la creencia incontestable de que el mundo se ve como se ve porque así es como es. Hay un pequeño paso de esta creencia a otra: «Otras personas verán el mundo de la misma manera que yo». Estas creencias, que han recibido el nombre de «realismo ingenuo», son esenciales para nuestro sentido de una realidad que compartimos con otras personas. Rara vez cuestionamos estas creencias.[2] Tenemos una única interpretación del mundo que nos rodea en un momento dado, y normalmente invertimos pocos esfuerzos en generar alternativas plausibles a ella. Una sola interpretación es suficiente, y la experimentamos como verdadera. No vamos por la vida imaginando formas alternativas de ver lo que vemos.

En el caso de los juicios profesionales, la creencia de que los demás ven el mundo de la misma manera que nosotros se refuerza cada día de múltiples maneras. En primer lugar, compartimos con nuestros colegas un lenguaje común y un conjunto de reglas sobre las consideraciones que

deben importar en nuestras decisiones. También tenemos la tranquilizadora experiencia de estar de acuerdo con otros en lo absurdo de los juicios que violan esas reglas. Experimentamos los desacuerdos ocasionales que encontramos con los colegas como fallos en sus juicios. Tenemos pocas oportunidades de advertir que las normas que nos hacen estar de acuerdo con otros son vagas, suficientes para eliminar algunas posibilidades, pero no para especificar una respuesta positiva compartida a un caso particular. Podemos vivir cómodamente con nuestros colegas sin notar que, de hecho, ellos no ven el mundo como nosotros.

Una agente de seguros que entrevistamos describió así su experiencia de veterana en su departamento: «Cuando era nueva, discutía el 75 por ciento de los casos con mi supervisor [...] Pasados unos años no necesitaba hacerlo —ahora se me considera una experta— [...] Con el tiempo, me volví cada vez más confiada en mi juicio». Como muchos de nosotros, esta persona había desarrollado la confianza en sus juicios principalmente ejerciendo su función juzgadora.

La psicología de este proceso es bien conocida. La confianza se nutre de la experiencia subjetiva de los juicios que se hacen con creciente fluidez y facilidad, en parte porque se asemejan a los juicios hechos en casos similares en el pasado. Con el tiempo, aquella agente aprendió a estar de acuerdo con su pasado, y la confianza en sus juicios aumentó. No dio ninguna indicación de que —después de la fase inicial de aprendizaje— hubiera aprendido a estar de acuerdo con los demás, de que hubiese comprobado hasta qué punto lo estaba o, incluso, de haber tratado de evitar que sus prácticas se alejaran de las de sus colegas.

En la compañía aseguradora, la ilusión de acuerdo se quebró solo por obra de la auditoría del ruido. ¿Cómo los líderes de la compañía de seguros no se habían dado cuenta de su problema con el ruido? Hay varias respuestas posibles, pero una que parece jugar un gran papel en muchos escenarios es simplemente la incomodidad del desacuerdo. La mayoría de las organizaciones valoran más el consenso y la armonía que la disensión y el conflicto. Los procedimientos establecidos suelen estar concebidos expresamente para reducir al mínimo la frecuencia de la exposición a desacuerdos reales, y, cuando esos desacuerdos se producen, hallarles una explicación.

Nathan Kuncel, profesor de Psicología en la Universidad de Minnesota e investigador destacado en la predicción del rendimiento, compartió con nosotros una historia que ilustra este problema. Kuncel estaba ayudando al departamento de admisiones de una escuela a revisar el proceso que siguen sus decisiones. Primeramente, una persona leía un expediente de solicitud, lo calificaba y lo entregaba con las calificaciones a un segundo lector que, a su vez, lo calificaba. Kuncel sugirió —por razones que resultarán obvias a lo largo de este libro— que sería preferible enmascarar las calificaciones del primer lector, para que no influyesen en las del segundo. La respuesta de la escuela fue: «Solíamos hacer eso, pero hubo tantos desacuerdos que cambiamos al sistema actual». Esta escuela no es la única organización en la que evitar el conflicto es un objetivo tan importante al menos como tomar la decisión acertada.

Consideremos otro mecanismo al que muchas empresas recurren: los escrutinios posteriores de juicios desafortunados. Estos escrutinios son útiles como mecanismos de aprendizaje. Sin embargo, si realmente se ha cometido un fallo —en el sentido de que un juicio se apartó de las normas profesionales—, discutirlo no será muy difícil. Los expertos llegarán fácilmente a la conclusión de que el juicio estaba muy alejado del consenso. (Podrán tomarlo como una rara excepción.) Un mal juicio es mucho más fácil de identificar que uno bueno. La denuncia de fallos clamorosos y la marginación de los colegas malos no ayudará a los profesionales a darse cuenta de hasta qué punto están en desacuerdo cuando hacen juicios que todos hallarán aceptables. Por el contrario, el fácil consenso sobre los malos juicios puede incluso reforzar la ilusión de acuerdo. La buena lección sobre la omnipresencia del ruido en el sistema jamás se aprenderá.

Esperamos que el lector empiece a compartir nuestra opinión de que el ruido del sistema es un problema serio. Su existencia no es una sorpresa; el ruido es una consecuencia de la naturaleza informal del juicio. Sin embargo, como veremos a lo largo de este libro, la cantidad de ruido que una organización observa cuando le echa una mirada seria casi siempre le provoca un *shock*. Nuestra conclusión es simple: dondequiera que haya juicio, hay ruido, y más del que se piensa.

SOBRE EL RUIDO DEL SISTEMA EN UNA COMPAÑÍA DE SEGUROS

- «Dependemos de la calidad de los juicios profesionales por parte de los agentes, los tasadores de reclamaciones y otros. Asignamos cada caso a un experto, pero operamos desde el supuesto erróneo de que otro experto produciría un juicio similar.»
- «El ruido del sistema es cinco veces mayor de lo que pensábamos —o de lo que podemos tolerar—. Sin una auditoría del ruido, nunca nos habríamos enterado. La auditoría del ruido rompió la ilusión de acuerdo.»
- «El ruido del sistema es un serio problema: nos cuesta cientos de millones.»
- «Dondequiera que haya juicio, hay ruido, y más del que pensamos.»

3
Decisiones singulares

Los casos que hemos analizado implican juicios que se hacen repetidamente. ¿Cuál es la sentencia correcta para alguien condenado por robo? ¿Cuál es la prima adecuada para un riesgo en particular? Aunque cada caso es único, se trata de «decisiones recurrentes». Los médicos que diagnostican a pacientes, los jueces que ven los casos de libertad bajo palabra, los responsables de las admisiones en una escuela que examinan solicitudes o los contables que rellenan los formularios de impuestos, son todos ejemplos de decisiones recurrentes.

Una auditoría del ruido como la que describimos en el capítulo anterior demuestra la existencia de ruido en las decisiones recurrentes. La variabilidad no deseada es fácil de definir y medir cuando profesionales intercambiables toman decisiones en casos similares. Parece mucho más difícil, si no imposible, aplicar la idea de «ruido» a una categoría de juicios que llamamos «decisiones singulares».

Consideremos, por ejemplo, la crisis a la que el mundo tuvo que enfrentarse en 2014. En África occidental, mucha gente moría a causa del ébola. Como el mundo estaba interconectado, las predicciones apuntaban a que las infecciones se propagarían rápidamente por todo el mundo y afectarían en particular a Europa y a Norteamérica. En Estados Unidos se hicieron llamamientos insistentes para que se cancelaran los vuelos procedentes de las regiones afectadas y para que se tomaran medidas tan drásticas como cerrar las fronteras. La presión política para avanzar en esa dirección era intensa, y personas eminentes y bien informadas querían tomar esas medidas.

El presidente Barack Obama se enfrentó a una de las decisiones más difíciles de su presidencia, una decisión a la que no se había enfrentado antes y a la que no volvió a enfrentarse. Decidió no cerrar ninguna frontera. En cambio, envió a tres mil personas —sanitarios y soldados— al África occidental. Dirigió una coalición internacional de diversas naciones, las

cuales no siempre trabajaban bien juntas, que utilizaría sus recursos y conocimientos para abordar el problema en su origen.

SINGULAR *VERSUS* RECURRENTE

Las decisiones que se toman solo una vez, como la respuesta del presidente al ébola, son singulares, porque no las toma de modo recurrente el mismo individuo o equipo, no tienen una respuesta preparada y vienen marcadas por características del todo únicas. Cuando se tuvo que hacer frente al ébola, el presidente Obama y su equipo no tenían precedentes reales en los que basarse. Las decisiones políticas importantes son con frecuencia buenos ejemplos de decisiones singulares, como también lo son la mayoría de las resoluciones críticas de los mandos militares.

Las decisiones que se toman en el ámbito privado, como optar por un trabajo, comprar una casa o proponer matrimonio, tienen las mismas características. Aunque no se trate del primer trabajo, la primera casa o el primer matrimonio, y a pesar de que incontables personas han tomado esas mismas decisiones con anterioridad, son decisiones experimentadas como actos únicos. En el mundo empresarial, a menudo se pide a los dirigentes de las empresas que tomen decisiones que parecen únicas: si se debe presentar una innovación que pueda cambiar las reglas del juego, durante cuánto tiempo echar el cierre en tiempos de pandemia, si se debe establecer una filial en un país extranjero o si se ha de capitular ante un gobierno que trata de regularlas.

Podría decirse que hay un continuo, no una diferencia de categoría, entre las decisiones singulares y las recurrentes. Los agentes de seguros pueden tratar algunos casos que les parecen muy fuera de lo común. Por el contrario, si alguien compra una casa por cuarta vez en su vida, es probable que empiece a pensar en la compra de una casa como una decisión recurrente. Sin embargo, los ejemplos extremos indican claramente que la diferencia es significativa. Ir a la guerra es una cosa; pasar por las revisiones anuales del presupuesto es otra.

EL RUIDO EN LAS DECISIONES SINGULARES

Las decisiones singulares se han tratado tradicionalmente como algo separado por completo de los juicios recurrentes que empleados intercambiables hacen de forma rutinaria en las grandes organizaciones. Mientras que los científicos sociales se han ocupado de las decisiones recurrentes, las decisiones singulares de alto riesgo han sido competencia de historiadores y gurús de las empresas. Los métodos de ambos tipos de decisiones han sido muy diferentes. En los análisis de las decisiones recurrentes, a menudo se ha adoptado un enfoque estadístico, ya que los científicos sociales evalúan muchas decisiones similares para discernir patrones, identificar regularidades y medir el grado de acierto. En cambio, las discusiones sobre decisiones singulares suelen adoptar un punto de vista causal; tienen carácter retrospectivo y se centran en la identificación de las causas de lo sucedido. Los análisis históricos, como los estudios de casos de éxitos y fracasos en la gestión, tienen por objeto entender cómo se hizo un juicio que era en esencia único.

La naturaleza de las decisiones singulares plantea una cuestión importante para el estudio del ruido. Hemos definido el «ruido» como una variabilidad no deseada en juicios sobre un mismo problema. Dado que los problemas singulares nunca se repiten exactamente, esta definición no es aplicable a ellos. Después de todo, la historia se refiere a algo que aconteció solo una vez. Nunca se podrá comparar la decisión de Obama de enviar sanitarios y soldados a África occidental en 2014 con las decisiones de otros presidentes estadounidenses sobre cómo resolver un problema particular en un momento particular (aunque aquí se puede especular). Podemos entender que se compare la decisión que una persona tome de casarse con alguien especial con las decisiones de otras personas similares, pero esa comparación no será tan relevante para quien tome dicha decisión como la comparación que hicimos entre las cuotas de los agentes de seguros para el mismo caso, porque él y su futuro cónyuge son únicos. No hay una forma directa de observar la presencia de ruido en las decisiones singulares.

Sin embargo, las decisiones singulares no están libres de los factores que producen ruido en las decisiones recurrentes. En el campo de tiro, los tiradores del equipo C (el equipo ruidoso) podrían haber ajustado la mira de su rifle en direcciones diferentes, o tal vez sus manos sean poco firmes. Si observamos solo al primer tirador del equipo, no tendríamos una idea de lo ruidoso que es el equipo, pero las fuentes de ruido están ahí. Del mismo

modo, quien toma una decisión singular imaginará que otro igual de competente que él, y con idénticos objetivos y valores, podría no extraer la misma conclusión de los mismos hechos. Y, como toda persona que toma una decisión, reconocerá que podría haber tomado una decisión diferente si algunos aspectos irrelevantes de la situación o del proceso de decidir hubieran sido diferentes.

En otras palabras, no podemos medir el ruido en una decisión singular, pero, si pensamos contrafácticamente, sabemos con certeza que el ruido existe. Así como la mano temblorosa del tirador implica que un solo disparo podría haber impactado en otro lugar, el ruido en los encargados de tomar decisiones y en el proceso de tomar una decisión implica que la decisión singular podría haber sido diferente.

Consideremos todos los factores que afectan a una decisión singular. Si los expertos encargados de analizar la amenaza del ébola y de preparar planes de respuesta hubieran sido otras personas con diferentes historiales y experiencias de la vida, ¿habría sido su propuesta al presidente Obama la misma? Si los mismos hechos se hubieran presentado de una manera algo diferente, ¿la conversación se habría desarrollado de la misma manera? Si el estado de ánimo de los principales actores hubiera sido distinto, o estos se hubieran reunido bajo una tormenta de nieve, ¿habría sido idéntica la decisión final? Vista bajo esta luz, la decisión singular no parece tan resuelta. Es plausible que, dependiendo de una serie de factores de los que ni siquiera somos conscientes, la decisión podría haber sido diferente.

Otro ejercicio de pensamiento contrafáctico sería este: consideremos cómo respondieron los distintos países y zonas a la crisis de la COVID-19. Aun cuando el virus los afectó, aproximadamente, al mismo tiempo y de manera similar, hubo grandes diferencias en la respuesta. Esta variación proporciona una clara evidencia de ruido en las decisiones que tomaron diferentes países. Pero ¿y si la epidemia hubiera afectado a un solo país? En este caso, no habríamos observado ninguna variabilidad. Sin embargo, nuestra incapacidad para observar alguna variabilidad no haría a la decisión menos ruidosa.

Esta discusión teórica es importante. Si las decisiones singulares son exactamente tan ruidosas como las recurrentes, las estrategias que reducen el ruido en las decisiones recurrentes deberían mejorar también la calidad de las decisiones singulares.

Esta parece una propuesta del todo contraria a la intuición. Cuando hemos de tomar una decisión única, es probable que el instinto nos haga verla como algo único. Incluso hay quienes afirman que las reglas del pensamiento probabilístico son del todo irrelevantes para las decisiones singulares tomadas en condiciones de incertidumbre, y que estas decisiones demandan un enfoque radicalmente diferente.

Nuestras observaciones sugieren lo contrario. Desde la perspectiva de la reducción del ruido, «una decisión singular es una decisión recurrente que ocurre solo una vez». Se tome una o cien veces una decisión, el objetivo debería ser hacerlo de manera que reduzca tanto el sesgo como el ruido. Y las prácticas que reducen el error tendrían que ser tan eficaces en nuestras decisiones únicas como en las repetidas.

SOBRE LAS DECISIONES SINGULARES

- «La forma de actuar ante esta oportunidad inusual te expone al ruido.»
- «Recuerda: una decisión singular es una decisión recurrente que se toma una sola vez.»
- «Las experiencias personales que te hicieron ser quien eres no son verdaderamente relevantes en esta decisión.»

SEGUNDA PARTE

Nuestra mente es un instrumento de medición

Una medición es, tanto en la vida cotidiana como en la ciencia, el acto de utilizar un instrumento para asignar un valor de una escala a un objeto. Medimos el largo de una alfombra en centímetros usando una cinta métrica. Medimos la temperatura en grados centígrados o Fahrenheit mediante un termómetro.

El acto de hacer un juicio es similar. Cuando los jueces determinan la pena de prisión apropiada para un delito, asignan un valor de una escala. También lo hacen los agentes de seguros cuando fijan un valor monetario para cubrir un riesgo, o los médicos cuando hacen un diagnóstico. (La escala no tiene por qué ser numérica: «culpable más allá de toda duda razonable», «melanoma avanzado» y «es conveniente la cirugía» son también juicios.)

El juicio puede entonces describirse como «medición en la que el instrumento es una mente humana». En la noción de «medición» va implícito el objetivo de la exactitud (acercarse a la verdad y minimizar el error). El objetivo del juicio no es impresionar ni tomar partido ni persuadir. Es importante señalar que el concepto de «juicio», tal y como aquí lo usamos, está tomado de la literatura psicológica técnica y es un concepto mucho más limitado que el de esa palabra en el lenguaje común. «Juzgar» no es sinónimo de «pensar», y «hacer juicios acertados» no es sinónimo de «tener buen juicio».

Un juicio así definido es una conclusión que puede resumirse en una palabra o una frase. Si un analista inteligente escribe un extenso informe que concluye que un régimen es inestable, solo esta conclusión es un juicio. El «juicio», como la «medición», se refiere tanto a la actividad mental de juzgar como a su producto. Y a veces usaremos la palabra «juzgador» como término técnico para referirnos a personas que emiten juicios que nada tienen que ver con la judicatura.

Aunque el objetivo es la exactitud, nunca se alcanza la perfección en el logro de este objetivo, ni siquiera en la medición científica, y mucho menos en el juicio. Siempre habrá errores, de los cuales unos se deberán al sesgo y otros al ruido.

Para experimentar cómo el ruido y el sesgo contribuyen al error, invitamos al lector a jugar a un juego que le llevará menos de un minuto. Si tiene un *smartphone* con cronómetro, probablemente tenga una función de «vuelta» que le permite medir intervalos de tiempo consecutivos sin detener el cronómetro o incluso sin mirar la pantalla. El objetivo es contar cinco vueltas consecutivas de exactamente diez segundos sin mirar el teléfono. Puede que quiera observar un intervalo de diez segundos unas cuantas veces antes de empezar. Adelante.

Ahora mire las duraciones de las vueltas grabadas en el teléfono. (El teléfono en sí mismo no está libre de ruido, pero hay muy poco de él.) Verá que las vueltas no son todas de diez segundos, y que varían en un rango sustancial. Intente reproducir exactamente el mismo cronometraje; no podrá hacerlo. La variabilidad que no ha podido controlar es un ejemplo de ruido.

Este hallazgo no es sorprendente, porque el ruido es universal en la fisiología y la psicología. La variabilidad entre los individuos es un dato biológico; no hay dos guisantes en una vaina que sean idénticos. Dentro de la misma persona, también hay variabilidad. Los latidos del corazón no son del todo regulares. No se puede repetir el mismo gesto con total precisión. Y, cuando un audiólogo nos examina la audición, habrá algunos sonidos tan suaves que nunca los oiremos, y otros tan fuertes que siempre los oiremos, pero también habrá algunos sonidos que unas veces oiremos y otras no.

Ahora mire las cinco medidas en su teléfono. ¿Ve un patrón? Por ejemplo, ¿duran las cinco vueltas menos de diez segundos, lo que sugiere que su «reloj» interno está corriendo rápido? El «sesgo» en esta simple tarea es la diferencia, positiva o negativa, entre la media de sus vueltas y diez segundos. El «ruido» es la variabilidad de los resultados, análoga a la dispersión de los disparos que vimos antes. En estadística, la medida más común de la variabilidad es la «desviación estándar», y es la que utilizaremos para medir el ruido en los juicios.[1]

Podemos concebir la mayoría de los juicios, específicamente los juicios predictivos, como similares a las mediciones que acabamos de hacer. Cuando hacemos una predicción, intentamos acercarnos al valor real. Un analista económico se propone acercarse todo cuanto puede al valor real del aumento del producto interior bruto (PIB) en el próximo año; un médico se propone hacer un diagnóstico correcto. (Nótese que el término «predicción», en el sentido técnico en que lo usamos en este libro, no

implica que se esté prediciendo el futuro: para nuestros propósitos, el diagnóstico de una patología existente es una predicción.)

Nos basaremos ampliamente en la analogía entre juicio y medición porque ayuda a explicar el papel del ruido en el error. Las personas que hacen juicios predictivos son como el tirador que apunta a la diana o el físico que se esfuerza por medir el peso real de una partícula. El ruido en sus juicios implica error. En pocas palabras, si un juicio da un valor real, dos juicios diferentes no pueden ser ambos correctos. Al igual que los instrumentos de medida, algunas personas generalmente muestran más errores que otras en una tarea en particular (tal vez debido a deficiencias en habilidades o de formación). Sin embargo, también como los instrumentos de medida, las personas que hacen juicios nunca son perfectas. Necesitamos comprender y medir sus errores.

La mayoría de los juicios profesionales son, por supuesto, mucho más complejos que la medición de un intervalo de tiempo. En el capítulo 4 definimos diferentes tipos de juicios profesionales y analizamos sus objetivos. En el capítulo 5 explicamos cómo medir el error y cómo cuantificar la contribución a este del ruido del sistema. El capítulo 6 profundiza en el ruido del sistema e identifica sus componentes, que son diferentes tipos de ruido. En el capítulo 7 analizamos uno de estos componentes: el ruido de ocasión. Por último, en el capítulo 8 mostramos cómo los grupos amplifican a menudo el ruido en los juicios.

De estos capítulos se infiere una sencilla conclusión: al igual que un instrumento de medición, la mente humana es imperfecta; en ella hay sesgo y ruido. ¿Por qué? ¿Y cuánto? Averigüémoslo.

4
Cuestiones de juicio

Este libro trata de los juicios profesionales entendidos en un sentido amplio, y parte de la base de que las personas que hacen tales juicios se proponen acertar. Sin embargo, el propio concepto de «juicio» implica un reconocimiento reticente de que nunca se puede estar seguro de que un juicio sea certero.

Consideremos las expresiones «cuestión de juicio» o «es un juicio personal». No consideramos que proposiciones como «el sol saldrá mañana» o «la fórmula del cloruro sódico es NaCl» sean juicios, porque se espera que las personas razonables estén totalmente de acuerdo con ellas. Una cuestión de juicio es una cuestión con cierta incertidumbre sobre lo que se juzga, en la que admitimos la posibilidad de que personas razonables y competentes puedan estar en desacuerdo.

Sin embargo, el desacuerdo admisible tiene un límite. De hecho, la palabra «juicio» se utiliza sobre todo cuando la gente cree que debe estar de acuerdo. Las cuestiones de juicio difieren de las cuestiones de opinión o de gusto en que las diferencias no resueltas son del todo aceptables. Los ejecutivos de la compañía de seguros que se sorprendieron ante el resultado de la auditoría del ruido no tendrían ningún problema si los tasadores de reclamaciones estuvieran en completo desacuerdo sobre si los Beatles eran mejores que los Rolling Stones, o si el salmón sabe mejor que el atún.

Las cuestiones de juicio, incluidas las profesionales, ocupan un espacio intermedio entre las cuestiones de hecho o de cálculo, por un lado, y las cuestiones de gusto u opinión, por otro. Están definidas por la «expectativa de un desacuerdo limitado».

El grado exacto de desacuerdo aceptable en un juicio es en sí mismo un juicio subjetivo que depende de la dificultad del problema. El acuerdo es especialmente fácil cuando un juicio es absurdo. Los jueces que difieren en gran medida en las sentencias que dictan en un caso de fraude común y corriente coincidirán en que tanto una multa de un dólar como una cadena perpetua no son razonables. Los jueces de los concursos de vino difieren

mucho en cuanto a los vinos que deberían recibir medallas, pero suelen ser unánimes en su desprecio por los rechazados.[1]

LA EXPERIENCIA DEL JUICIO: UN EJEMPLO

Antes de seguir hablando de la experiencia del juicio, pedimos al lector que él mismo haga un juicio. Asimilará mejor el resto de este capítulo si hace este ejercicio hasta el final.

Imagine que es miembro de un equipo encargado de evaluar candidatos para el puesto de director ejecutivo en una empresa financiera regional de éxito moderado que se enfrenta a una competencia cada vez mayor. Se le pide que evalúe la probabilidad de que un candidato tenga éxito tras dos años en el puesto (el éxito se define aquí simplemente como el del candidato que ha mantenido su puesto de director ejecutivo pasados dos años) y que exprese la probabilidad en una escala de 0 (imposible) a 100 (seguro).

Michael Gambardi tiene treinta y siete años. Ha ocupado varios puestos desde que se graduó en la Harvard Business School hace doce años. Empezó muy pronto como fundador e inversor de dos *start-ups* que fracasaron sin atraer mucho apoyo financiero. Luego se unió a una gran compañía de seguros, y rápidamente ascendió al puesto de director de operaciones para Europa. En ese puesto, inició y gestionó una importante mejora en la resolución puntual de reclamaciones. Sus colegas y subordinados lo describieron como eficaz, pero también como un hombre dominante y desabrido, y hubo un gran reemplazo de ejecutivos durante su desempeño. Sus colegas y subordinados también dieron fe de su integridad y de su disposición a asumir la responsabilidad de los fracasos. Durante los últimos dos años ha sido director ejecutivo de una compañía financiera de tamaño medio que, cuando él entró, corría el riesgo de fracasar. Estabilizó la compañía, donde se lo considera muy capaz aunque difícil para trabajar con él. Ha manifestado su interés en continuar. Los especialistas en recursos humanos que lo entrevistaron hace unos años le dieron calificaciones altas en cuanto a creatividad y energía, pero también lo describieron como arrogante y, a veces, tiránico.

Recuerde el lector que Michael es candidato a un puesto de director ejecutivo en una firma financiera regional de éxito moderado que se enfrenta a una competencia cada vez mayor. ¿Cuál es la probabilidad de que Michael, si es contratado, siga en su puesto pasados dos años? Elija un número específico en el rango de 0 a 100 antes de seguir leyendo. Vuelva a leer la descripción si lo necesita.

Si se ha tomado en serio la tarea, es probable que la haya encontrado difícil. Hay una gran cantidad de información, buena parte de ella, según parece, inconsistente. Habrá tenido que hacer esfuerzos por formarse la idea coherente que necesitaba para emitir un juicio. En el proceso de formación de esa idea, se habrá centrado en algunos detalles que parecían importantes, y muy probablemente habrá ignorado otros. Si le pedimos que explique su elección de un número, seguro que mencionará algunos hechos destacados, pero no los suficientes para una justificación completa de su juicio.

El proceso que siguió su pensamiento ilustra tres características de la operación mental que llamamos «juicio»:

- De todas las pistas proporcionadas por la descripción (que solo son un subconjunto de lo que necesitaría saber), se habrá fijado en unas más que en otras, sin ser del todo consciente de las elecciones que hizo. ¿Se ha dado cuenta de que Gambardi es un nombre italiano? ¿Recuerda la escuela de negocios a la que asistió? Este ejercicio fue diseñado para sobrecargarle y que no pudiera abarcar todos los detalles del caso. Es muy probable que su recuerdo de lo que le presentamos sea diferente del de otros lectores. La atención selectiva y el recuerdo selectivo son una fuente de variabilidad entre las personas.
- A continuación, integró informalmente estas pistas en una impresión general de las perspectivas de Gambardi. La palabra clave es aquí «informalmente». No construyó un plan para responder a la pregunta. Sin ser del todo consciente de lo que estaba haciendo, su mente trabajó para formarse una idea coherente de los puntos fuertes y débiles de Michael y de los desafíos a los que había tenido que hacer frente. La informalidad le permitió trabajar con rapidez. Esto también produce variabilidad: un proceso formal como la suma de una columna de números garantiza resultados idénticos, pero en una operación informal es inevitable que haya algo de ruido.
- Finalmente, convirtió esta impresión general en un número de una «escala» de probabilidad de éxito. Hacer coincidir un número entre 0 y 100 con una impresión es un proceso misterioso sobre el que volveremos en el capítulo 14. Una vez más, el lector no sabrá exactamente por qué respondió como lo hizo. ¿Por qué eligió, digamos, 65 y no 61 o 69? Lo más probable es que, en algún momento, se le haya ocurrido un número. Comprobó si ese número le parecía correcto; y, si no, se le ocurrió otro número. Esta parte del proceso también es una fuente de variabilidad entre las personas.

Dado que cada uno de estos tres pasos en un proceso complejo de formación de un juicio implica cierta variabilidad, no tendría que sorprendernos encontrar mucho ruido en las respuestas sobre Michael Gambardi. Si el lector pide a unos cuantos amigos que lean el caso, es casi seguro que encontrará que sus estimaciones de la probabilidad de éxito en Gambardi están muy dispersas. Cuando mostramos el caso a ciento quince estudiantes de MBA, sus respuestas oscilaron entre el 10 por ciento y el 95 por ciento. Eso es muchísimo ruido.

Por cierto, el lector habrá notado que el ejercicio del cronómetro y el problema de Gambardi ilustran dos tipos de ruido. La variabilidad de los juicios sobre sucesivas pruebas con el cronómetro es ruido «dentro» de un solo juzgador (uno mismo), mientras que la variabilidad de los juicios en el caso de Gambardi es ruido entre diferentes juzgadores. En términos de medición, el primer problema ilustra la fiabilidad «intrapersonal», y el segundo ilustra la fiabilidad «interpersonal».

LO QUE EL JUICIO ASPIRA A CONSEGUIR: LA SEÑAL INTERNA

La respuesta a la pregunta sobre Gambardi es un «juicio predictivo», tal y como hemos definido el término. Sin embargo, difiere en aspectos importantes de otros juicios que llamamos «predictivos», como la temperatura máxima de mañana en Bangkok, el resultado del partido de fútbol de esta noche o el de las próximas elecciones presidenciales. Si uno no está de acuerdo con un amigo sobre estos asuntos, en algún momento descubrirá quién tenía razón. Sin embargo, si el lector no está de acuerdo con su amigo sobre Gambardi, el tiempo «no» dirá quién había acertado, y eso es por una sencilla razón: Gambardi no existe.

Aunque la pregunta se refiriera a una persona real y conociéramos el resultado, no se puede confirmar ni desconfirmar un único juicio de probabilidad (distinto del 0 o del 100 por ciento). El resultado no revela cuál era la probabilidad *ex ante*. Si un acontecimiento al que se le asignó una probabilidad del 90 por ciento no se produce, el juicio de probabilidad no fue necesariamente malo. Después de todo, los resultados que solo tienen una probabilidad del 10 por ciento de producirse acaban ocurriendo el 10 por ciento de las veces. El ejercicio de Gambardi es un ejemplo de

juicio predictivo «no verificable» por dos razones distintas: Gambardi es ficticio y la respuesta es probabilística.

Muchos juicios profesionales no son verificables. Salvo errores garrafales, los agentes de seguros nunca sabrán, por ejemplo, si una póliza concreta era demasiado elevada o demasiado baja. Otras predicciones pueden no ser verificables porque son condicionales. «Si vamos a la guerra, nos aplastarán» es una predicción importante, pero es probable que no se ponga a prueba (esperamos). También puede que las previsiones sean a demasiado largo plazo para que los profesionales que las hacen tengan que rendir cuentas (como, por ejemplo, una estimación de las temperaturas medias a finales del siglo XXI).

¿Cambió la naturaleza no verificable del ejercicio de Gambardi la forma de abordarlo por parte del lector? ¿Se preguntó este, por ejemplo, si Gambardi era real o ficticio? ¿Se preguntó si el resultado se iba a revelar más adelante en el texto? ¿Reflexionó sobre el hecho de que, aunque así fuera, la revelación no le daría la respuesta a la pregunta que se le hacía? Probablemente no, porque estas consideraciones no le parecerían relevantes cuando respondió a la pregunta.

La verificabilidad no cambia la experiencia del juicio. Hasta cierto punto, tal vez el lector pensase más en un problema cuya respuesta se le revelaría pronto, porque el temor a quedar en evidencia hace que la mente se concentre. Y, a la inversa, pudo negarse a pensar mucho en un problema tan hipotético que llegó a encontrarlo absurdo, («Si Gambardi tuviera tres piernas y pudiera volar, ¿sería un mejor director general?») Sin embargo, por lo común, un problema hipotético plausible se aborda de la misma manera que uno real. Esta similitud es importante para la investigación psicológica, que en gran parte se vale de problemas inventados.

Dado que no hay un resultado —y puede que el lector ni siquiera se haya preguntado si lo habrá—, no estaba tratando de minimizar el error relativo a ese resultado. El lector intentó acertar en el juicio, dar con un número en el que tuviera la suficiente confianza para presentarlo como su respuesta. Por supuesto, no estaba totalmente seguro de esa respuesta como sí lo estaría de que 6 por 4 son 24. Era consciente de cierta incertidumbre (y, como veremos, tal vez haya más de ella de lo que reconoce); pero, en algún

momento, decidió que ya no estaba progresando y decidió dar una respuesta.

¿Qué es lo que le hizo sentir que había acertado en el juicio o, al menos, que había acertado bastante como para que esa fuera su respuesta? Sugerimos que esta sensación es una «señal interna de cumplimiento del juicio» sin relación con ninguna información externa. Habría tenido la sensación de que su respuesta era correcta de haberle parecido que casaba lo suficientemente bien con las evidencias. Sin embargo, una respuesta de 0 o 100 no le daría esa sensación de ajuste: la confianza que implica no es coherente con unas evidencias desordenadas, ambiguas y conflictivas como las aportadas. Pero el número por el que se había decidido, sea cual fuere, le daba la sensación de coherencia que necesitaba. El objetivo del juicio, tal y como lo experimentó, era encontrar una solución coherente.

La característica esencial de esta señal interna es que la sensación de coherencia es parte de la experiencia del juicio. No está supeditada a un resultado real. En consecuencia, la señal interna está disponible tanto para los juicios no verificables como para los reales, verificables. Esto explica por qué emitir un juicio sobre un personaje ficticio como Gambardi se parece mucho a emitir un juicio sobre el mundo real.

CÓMO SE EVALÚA EL JUICIO: EL RESULTADO Y EL PROCESO

La verificabilidad no cambia la experiencia del juicio tal y como este acontece, pero cambia la evaluación del juicio después del hecho.

Los juicios verificables pueden ser puntuados por un observador objetivo con una simple medida de error: la diferencia entre el juicio y el resultado. Si un meteorólogo ha dicho que la temperatura máxima de hoy sería de 21 grados centígrados y es de 18 grados centígrados, el meteorólogo ha cometido un error de 3 grados. Evidentemente, este enfoque no funciona para juicios no verificables, como el problema de Gambardi, que no tienen un resultado real. ¿Cómo vamos entonces a decidir lo que constituye un buen juicio?

La respuesta es que hay una segunda forma de evaluar los juicios. Este enfoque se aplica tanto a los verificables como a los no verificables. Consiste en evaluar el «proceso» del juicio. Cuando hablamos de juicios buenos o malos, podemos referirnos tanto al resultado (por ejemplo, el

número que se dio en el caso de Gambardi) como al proceso (lo que el lector hizo para llegar a ese número).

Un método para evaluar el proceso del juicio es observar cómo acontece ese proceso cuando se aplica a un gran número de casos. Consideremos, por ejemplo, que un analista político ha asignado probabilidades de ganar a un gran número de candidatos en unas elecciones locales. Otorgó a cien de estos candidatos un 70 por ciento de posibilidades de ganar. Si setenta de ellos son finalmente elegidos, tenemos un buen indicio de la habilidad del analista en el uso de la escala de probabilidad. Los juicios son verificables como un conjunto, aunque ningún juicio de probabilidad individual puede declararse correcto o equivocado. Del mismo modo, el sesgo a favor o en contra de un grupo concreto puede establecerse mejor examinando los resultados estadísticos de un número sustancial de casos.

Otra cuestión que puede plantearse sobre el proceso del juicio es si este se ajusta a los principios de la lógica o a la teoría de la probabilidad. Un gran número de investigaciones sobre los sesgos cognitivos del juicio han ido en esta línea.

Centrarse en el proceso del juicio, más que en su resultado, permite evaluar la calidad de los juicios que no son verificables, como los juicios sobre problemas ficticios o las previsiones a largo plazo. No podemos compararlos con un resultado conocido, pero sí saber si se han hecho de modo incorrecto. Y, cuando pasemos a la cuestión de «mejorar» los juicios en lugar de solo evaluarlos, nos centraremos también en el proceso. Todos los procedimientos que recomendamos en este libro para reducir el sesgo y el ruido tienen como objetivo adoptar el proceso que minimice el error sobre un conjunto de casos similares.

Hemos contrapuesto dos formas de evaluar un juicio: comparándolo con un «resultado» y valorando la calidad del «proceso» que ha conducido a él. Obsérvese que, cuando el juicio es verificable, las dos formas de evaluarlo pueden llegar a conclusiones diferentes sobre un mismo caso. Un analista hábil y cuidadoso que utilice las mejores herramientas y técnicas posibles a menudo no acertará con la cifra correcta al hacer una previsión de inflación trimestral. Mientras tanto, en un solo trimestre, un chimpancé lanzador de dardos a veces acertará.

Los estudiosos de la toma de decisiones ofrecen un consejo claro para resolver esta tensión: centrarse en el proceso, no en el resultado de un solo

caso. Sin embargo, reconocemos que esta no es una práctica habitual en la vida real. Los profesionales suelen ser evaluados en función del grado de coincidencia de sus juicios con los resultados verificables, y, si se les pregunta a qué aspiran con sus juicios, lo que responderán es que a una buena aproximación.

En resumen, lo que suele perseguirse en los juicios verificables es una predicción que coincida con el resultado. Lo que realmente se intenta tener, independientemente de la verificabilidad, es la señal interna de cumplimiento que proporciona la coherencia entre los hechos del caso y la sentencia. Y, lo que se debería intentar, normativamente hablando, es determinar el proceso que produzca el mejor juicio sobre un conjunto de casos similares.

JUICIOS EVALUATIVOS

En este capítulo nos hemos centrado hasta ahora en tareas del juicio predictivo, y la mayoría de los juicios que discutiremos son de ese tipo. Sin embargo, en el capítulo 1, que trata del juez Frankel y el ruido en las sentencias de los jueces federales, hemos examinado otro tipo de juicio. Condenar a un malhechor no es una predicción. Es un «juicio evaluativo» que busca adecuar la pena a la gravedad del delito. Los jueces de un concurso de vino o los críticos gastronómicos emiten juicios evaluativos. Los profesores que califican redacciones, los jueces de competiciones de patinaje sobre hielo y los comités que conceden subvenciones a proyectos de investigación emiten juicios evaluativos.

En las decisiones que implican múltiples opciones y compensaciones entre ellas se hace un tipo diferente de juicio evaluativo. Pensemos en los directivos que seleccionan entre los candidatos a una contratación, en los equipos de gestión que deben decidir opciones estratégicas o incluso en los presidentes que deciden cómo responder a una epidemia en África. Sin duda, todas estas decisiones se basan en juicios predictivos que proporcionan *inputs*; por ejemplo, el rendimiento de una candidata en su primer año, la respuesta del mercado de valores a un determinado movimiento estratégico o la rapidez con que se propagará la epidemia si no se controla. Sin embargo, las decisiones finales implican un equilibrio entre

los pros y los contras de varias opciones, y estos equilibrios se consiguen por medio de juicios evaluativos.[2]

Al igual que los juicios predictivos, los juicios evaluativos implican una expectativa de desacuerdo limitado. Ningún juez federal que se precie dirá: «Este es el castigo que más me gusta, y no me importa en absoluto que mis colegas piensen de otra manera». Y los responsables de la toma de decisiones que escogen entre varias opciones estratégicas esperan que los colegas y observadores que disponen de la misma información y comparten los mismos objetivos estén de acuerdo con ellos o, al menos, no discrepen demasiado. Los juicios evaluativos dependen, en parte, de los valores y preferencias de quienes los hacen, pero no son meras cuestiones de gusto o de opinión.

Por esta razón, la frontera entre los juicios predictivos y los evaluativos es difusa, y las personas que los emiten a menudo no son conscientes de ello. Los jueces que dictan las sentencias o los profesores que califican redacciones piensan mucho en su tarea y se esfuerzan por encontrar la respuesta «correcta». Desarrollan la confianza en sus juicios y en las justificaciones que tienen para ellos. Los profesionales sienten más o menos lo mismo, actúan más o menos del mismo modo y dicen más o menos lo mismo para justificarse cuando sus juicios son predictivos («¿Cómo se venderá este nuevo producto?») y cuando son evaluativos («¿Cómo se ha desempeñado mi asistente este año?»).

LAS CONSECUENCIAS NEGATIVAS DEL RUIDO

La observación de ruido en los juicios predictivos siempre indica que algo va mal. Si dos médicos no se ponen de acuerdo en un diagnóstico o dos analistas no se ponen de acuerdo sobre las ventas del próximo trimestre, al menos uno de ellos ha de estar equivocado. El error puede deberse a que uno de ellos es menos perspicaz y, por tanto, más propenso a equivocarse, o a alguna otra fuente de ruido. Independientemente de la causa, el juicio incorrecto puede tener serias consecuencias para quienes confían en los diagnósticos y pronósticos de estas personas.

El ruido en los juicios evaluativos es problemático por una razón diferente. En cualquier sistema en el que se supone que los juzgadores son intercambiables y se asignan de forma casi aleatoria, los grandes

desacuerdos sobre el mismo caso quebrantan las expectativas de justicia y coherencia. Si hay grandes diferencias en las condenas impuestas a un mismo tipo de acusado, estamos en el ámbito de las «crueldades arbitrarias» que denunció el juez Frankel. Incluso los jueces que creen en el valor de la individualización de las sentencias y que no están de acuerdo en la condena impuesta a un ladrón, estarán de acuerdo en que un nivel de desacuerdo que convierta un juicio en una lotería es problemático. Lo mismo sucede (aunque de forma menos dramática) cuando se dan calificaciones muy diferentes a un mismo ejercicio de redacción, distintas calificaciones relativas a la seguridad alimentaria a un mismo restaurante o diferentes puntuaciones a un mismo patinador sobre hielo; o cuando una persona que sufre de depresión percibe prestaciones de la Seguridad Social por incapacidad, mientras que otra persona con la misma enfermedad no recibe nada.

Incluso cuando la injusticia es solo una preocupación menor, el ruido del sistema plantea otro problema. Las personas que se ven afectadas por juicios evaluativos esperan que los valores que reflejan esos juicios sean los del sistema, no los de los juzgadores individuales. Algo va mal si un cliente que se queja de un ordenador portátil defectuoso recibe el reembolso íntegro y otro obtiene una mera disculpa; o si un empleado que lleva cinco años en una empresa pide un ascenso y lo consigue, mientras que la solicitud de otro empleado, cuyo rendimiento es por lo demás idéntico, es rechazada de forma educada. El ruido del sistema es inconsistencia, y esta daña la credibilidad del sistema.

INDESEADO, PERO MENSURABLE

Todo lo que necesitamos para medir el ruido son múltiples juicios sobre el mismo problema. No necesitamos conocer un valor real. Como ilustra la historia del campo de tiro en la introducción, cuando miramos el dorso del blanco la diana es invisible, pero podemos ver la dispersión de los disparos. Con solo saber que todos los tiradores apuntaban a la misma diana, podemos medir el ruido. Esto es lo que hace una auditoría del ruido. Si pedimos a todos nuestros analistas que estimen las ventas del próximo trimestre, la dispersión de sus pronósticos es ruido.

Esta diferencia entre el sesgo y el ruido es esencial para el propósito práctico de mejorar los juicios. Acaso se encuentre paradójico afirmar que podemos mejorar los juicios cuando no podemos verificar si son correctos. Sin embargo, podemos hacerlo (si empezamos midiendo el ruido). Independientemente de que el objetivo del juicio sea la exactitud o un equilibrio más complejo entre valores, el ruido es indeseado, y con frecuencia medible. Y, una vez medido el ruido, es posible, en general, reducirlo, como veremos en la quinta parte de este libro.

SOBRE EL JUICIO PROFESIONAL

- «Esto es una cuestión de juicio. No se puede esperar que las personas estén totalmente de acuerdo.»
- «Sí, es una cuestión de juicio, pero algunos juicios son tan excéntricos que son erróneos.»
- «Tu elección entre los candidatos fue solo una expresión de tus gustos, no un juicio serio.» ■ «Una decisión requiere tanto juicios predictivos como evaluativos.»

5
La medición del error

Es obvio que un sesgo consistente puede producir costosos errores. Si una báscula añade una cantidad constante a un peso, si un gestor entusiasta predice que los proyectos tardarán la mitad del tiempo que acaban tardando o si un ejecutivo tímido es excesivamente pesimista sobre las ventas futuras año tras año, el resultado serán numerosos y graves errores.

Ahora hemos visto que el ruido también puede producir costosos errores. Si la mayoría de las veces un gestor predice que los proyectos tardarán la mitad del tiempo que finalmente necesitaron, y otras predice que tardarán el doble del tiempo que realmente necesitaron, no sirve de nada decir que el gestor está «de promedio» en lo cierto. Los errores diferentes se suman; no se cancelan.

Por tanto, una cuestión importante es saber cómo y cuánto contribuyen el sesgo y el ruido al error. Este capítulo tratará de responder a esta pregunta. Su mensaje básico es sencillo: en todo tipo de juicios profesionales, siempre que el objetivo sea acertar, «el sesgo y el ruido juegan el mismo papel en el cálculo del error global». En unos casos, lo que más contribuya al error será el sesgo, y en otros, el ruido (y estos casos son más comunes de lo que se esperaría). Sin embargo, en todos los casos, disminuir el ruido tiene el mismo efecto en el error global que reducir el sesgo en la misma cantidad. Por ello, la medición y la reducción del ruido deben tener la misma prioridad que la medición y la reducción del sesgo.

Esta conclusión se basa en un método particular de medición del error que tiene una larga historia y que es generalmente aceptado en la ciencia y en la estadística. En este capítulo ofrecemos una visión general introductoria de esa historia y un esbozo del razonamiento subyacente.

¿DEBE GOODSELL REDUCIR EL RUIDO?

Imaginemos, para empezar, una gran empresa de venta al por menor llamada GoodSell que emplea a muchos analistas de ventas, cuyo trabajo

consiste en predecir la cuota de mercado de GoodSell en varias zonas. Quizá Amy Simkin, jefa del departamento de previsión de GoodSell, tras leer un libro sobre el tema del ruido, haya realizado una auditoría sobre este. Para ello, todos los analistas han hecho estimaciones independientes de la cuota de mercado en una misma zona.

La figura 3 muestra los resultados (inverosímilmente regulares) de la auditoría del ruido. Amy puede ver que los pronósticos se distribuyen en la conocida curva en forma de campana, también conocida como «distribución normal o gaussiana».

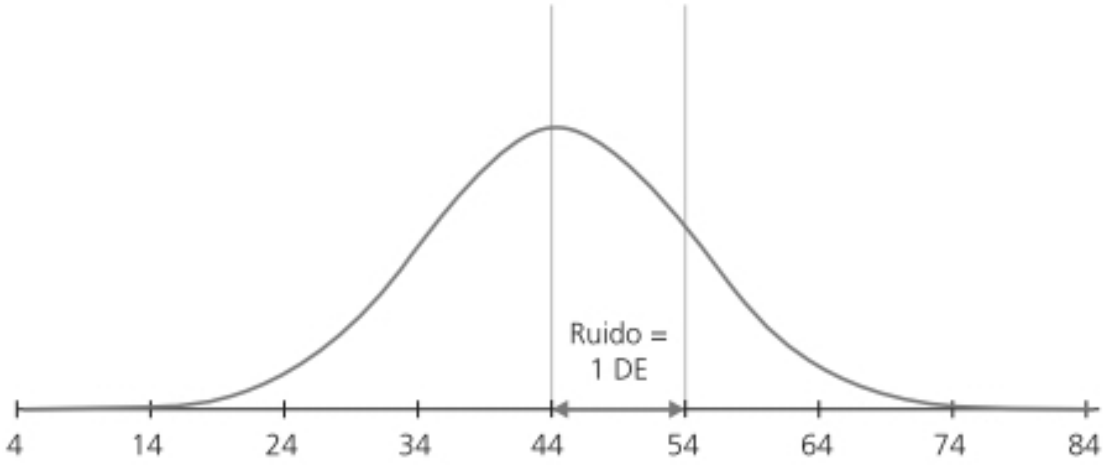

FIGURA 3. Distribución de las predicciones de la cuota de mercado de GoodSell para una zona

La predicción más frecuente, representada por la cima de la campana, es del 44 por ciento. Amy también puede ver que el sistema de previsión de la empresa es muy ruidoso: los pronósticos, que serían idénticos si todos fueran exactos, varían en un rango considerable.

Podemos asignar un número a la cantidad de ruido en el sistema de previsión de GoodSell. Tal y como en el uso del cronómetro para medir las vueltas, podemos calcular la «desviación estándar» de los pronósticos. Como su nombre indica, la desviación estándar representa una distancia típica de la media. En este ejemplo, son diez puntos porcentuales. Como ocurre en toda distribución normal, aproximadamente dos tercios de los pronósticos están contenidos dentro de una desviación estándar a cada lado de la media (en este ejemplo, entre un 34 por ciento y un 54 por ciento de la

cuota de mercado). Amy tiene ahora una estimación de la cantidad de ruido del sistema en los pronósticos de la cuota de mercado. (Una mejor auditoría del ruido contaría con varios problemas de previsión para obtener una estimación más sólida, pero uno solo es suficiente para nuestro propósito aquí.)

Como en el caso de los ejecutivos de la compañía de seguros real del capítulo 2, Amy está sorprendida por los resultados y quiere tomar medidas. La cantidad inaceptable de ruido indica que los analistas no son disciplinados en la implementación de los procedimientos que se espera que sigan. Amy pide autorización para contratar a un consultor en materia de ruido para lograr una mayor uniformidad y disciplina en el trabajo de sus analistas. Por desgracia, no consigue la autorización. La respuesta de su jefe parece bastante sensata: ¿cómo podríamos reducir los errores si no sabemos si nuestros pronósticos son o no acertados? Si hay un gran error medio en los pronósticos (es decir, un gran sesgo), sin duda lo prioritario sería estudiarlo. Antes de emprender algo para mejorar sus pronósticos, concluye, GoodSell debe esperar hasta saber si son acertados.

Un año después de la primera auditoría, se conoce el resultado que los analistas intentaban predecir. La cuota de mercado en la zona resultó ser del 34 por ciento. Ahora conocemos también el error de cada analista, que es simplemente la diferencia entre el pronóstico y el resultado. El error es 0 para un pronóstico del 34 por ciento, del 10 por ciento para el pronóstico medio del 44 por ciento y del -10 por ciento para un pronóstico del 24 por ciento.

La figura 4 muestra la distribución de los errores. Es la misma que la distribución de los pronósticos de la figura 3, pero se ha restado el valor real (34 por ciento) de cada predicción. La forma de la distribución no ha cambiado, y la desviación estándar (nuestra medida del ruido) sigue siendo del 10 por ciento.

La diferencia entre las figuras 3 y 4 es semejante a la diferencia entre un patrón de disparos visto por la cara y por el dorso del blanco en las figuras 1 y 2 (véase la introducción). Conocer la posición del objetivo no era necesario para observar el ruido en los disparos; del mismo modo, conocer el resultado real no añade nada a lo que ya se sabía sobre el ruido en los pronósticos.

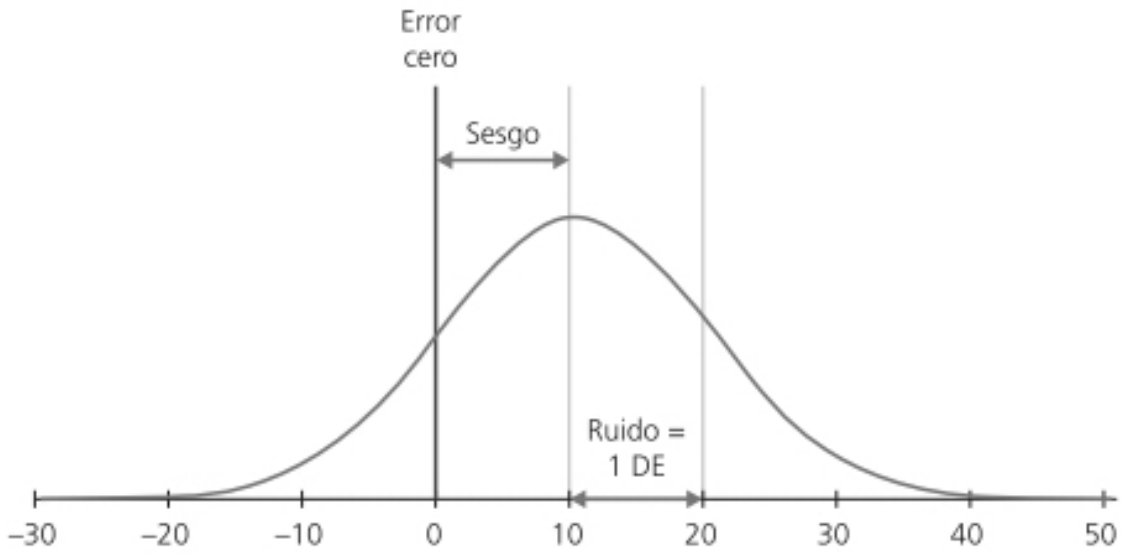

FIGURA 4. Distribución de los errores en los pronósticos de GoodSell para una zona

Amy Simkin y su jefe saben ahora algo que antes desconocían: la cantidad de sesgo en los pronósticos. El sesgo es simplemente la media de los errores, que en este caso también es del 10 por ciento. El sesgo y el ruido resultan ser numéricamente idénticos en este conjunto de datos. (Para ser claros, esta igualdad de ruido y sesgo no es en absoluto una regla general, pero un caso en el que el sesgo y el ruido son iguales facilita la comprensión de sus papeles.) Podemos ver que la mayoría de los analistas cometieron un error optimista, es decir, sobrestimaron la cuota de mercado que se alcanzaría: la mayoría de ellos se equivocó en el lado derecho de la línea vertical de error cero. (De hecho, utilizando las propiedades de la distribución normal, sabemos que este es el caso en el 84 por ciento de las predicciones.)

Como observa el jefe de Amy con una satisfacción apenas disimulada, él tenía razón. ¡Había mucho sesgo en las predicciones! Y ahora es evidente que reducir el sesgo sería algo bueno. Pero, se pregunta Amy, ¿habría sido una buena idea hace un año —y sería una buena idea ahora— reducir también el ruido? ¿Cómo se compararía el valor de esta mejora con el de la reducción del sesgo?

Para responder a la pregunta de Amy, necesitamos una «regla de puntuación» para los errores, una forma de ponderar y combinar los errores individuales en una única medida de error global. Afortunadamente, esta herramienta existe. Se trata del «método de los mínimos cuadrados», inventado en 1795 por Carl Friedrich Gauss, célebre prodigio de las matemáticas nacido en 1777, que inició una carrera de grandes descubrimientos ya en su adolescencia.[1]

Gauss propuso una regla para puntuar la contribución de los errores individuales al error global. Su medida de error global —llamado «error cuadrático medio» (ECM)— es la media de los cuadrados de los errores individuales de medición.

Los argumentos detallados de Gauss para su método de la medición del error global están fuera del alcance de este libro, y su solución no aparece a simple vista. ¿Por qué utilizar los cuadrados de los errores? La idea parece arbitraria, incluso extraña. Sin embargo, se basa, como se verá, en una intuición que es casi seguro que cualquiera compartirá.

Para ver por qué, pasemos a lo que parece un problema completamente diferente, pero que resulta ser el mismo. Imaginemos que nos dan una regla y nos piden que midamos la longitud de una línea con una precisión milimétrica. Se nos permite hacer cinco mediciones. Los triángulos que apuntan hacia abajo en la línea de la figura 5 representan estas mediciones.

Como puede verse, las cinco mediciones están todas entre 971 y 980 milímetros. ¿Cuál es la mejor estimación de la longitud real de la línea? Hay dos aspirantes obvias. Una posibilidad es la del número mediano, la medida que se sitúa entre las dos medidas más cortas y las dos más largas. Es la de 973 milímetros. La otra posibilidad es la media aritmética, conocida en el lenguaje común como el «promedio», que en este ejemplo es de 975 milímetros y se muestra con una flecha que apunta hacia arriba. La intuición probablemente favorezca la media, y esta intuición es correcta. La media contiene más información; está condicionada por la magnitud de los números, mientras que la mediana solo lo está por su orden.

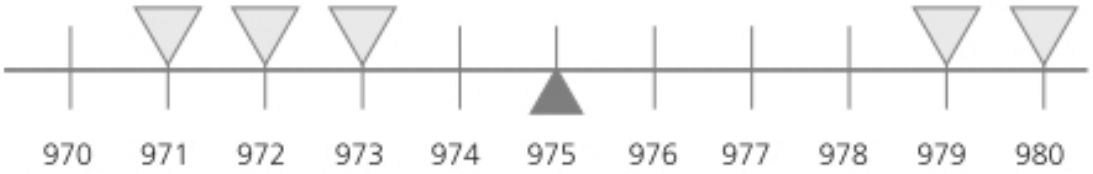

FIGURA 5. Cinco mediciones de la misma longitud

Hay un estrecho vínculo entre este problema de estimación, sobre el que se tiene una clara intuición, y el de la medición del error global que aquí nos ocupa. Son, de hecho, dos caras de la misma moneda. Esto se debe a que la mejor estimación es la que minimiza el error global de las mediciones disponibles. Por tanto, si la intuición de que la media es la mejor estimación es correcta, la fórmula que se utilice para medir el error global debe ser una fórmula que produzca la media aritmética como el valor con el que se minimiza el error.

El ECM tiene esta propiedad, y, de hecho, es la única definición de error global que la tiene. En la figura 6, hemos calculado el valor del ECM en el conjunto de cinco mediciones para diez posibles valores enteros de la longitud real de la línea. Por ejemplo, si el valor real fuera 971, los errores en las cinco mediciones serían 0, 1, 2, 8 y 9. Los cuadrados de estos errores suman 150, y su media es 30. Se trata de una cifra elevada, que refleja el hecho de que la estimación está lejos del valor real. Puede verse que el ECM decrece a medida que nos acercamos a 975 —la media—, y vuelve a crecer más allá de ese punto. La media es nuestra mejor estimación, porque es el valor que minimiza el error global.

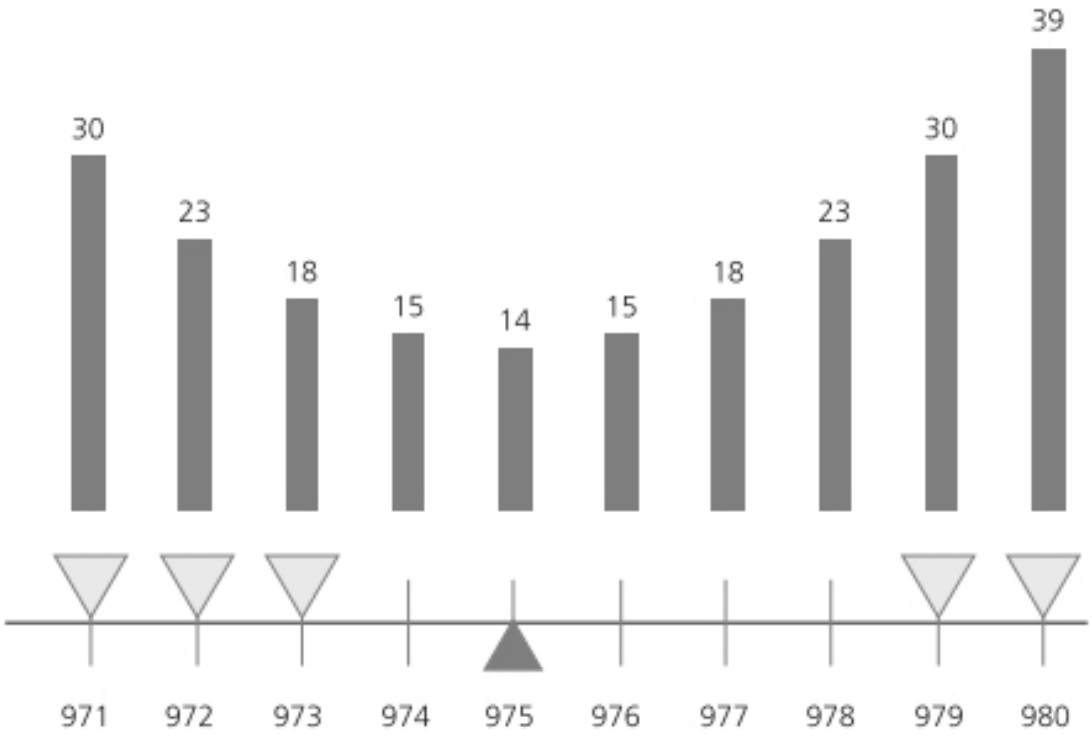

FIGURA 6. Error cuadrático medio (ECM) para diez valores posibles de la longitud real

También puede verse que el error global aumenta rápidamente cuando su estimación diverge de la media. Cuando la estimación aumenta en solo 3 milímetros, de 976 a 979, por ejemplo, el ECM se duplica. Esta es una característica clave del ECM: la elevación al cuadrado da un peso mucho mayor a los errores grandes que a los pequeños.

Ahora se ve por qué la fórmula de Gauss para medir el error global se llama «error cuadrático medio» y por qué su método de la estimación se llama «método de los mínimos cuadrados». El cuadrado de los errores es su idea central, y cualquier otra fórmula no sería compatible con la intuición de que la media es la mejor estimación.

Las ventajas del método de Gauss fueron rápidamente reconocidas por otros matemáticos. Entre sus muchas hazañas, Gauss utilizó el ECM (y otras innovaciones matemáticas) para resolver un rompecabezas que había derrotado a los mejores astrónomos de Europa: el redescubrimiento de Ceres, un asteroide que había sido solo brevemente rastreado antes de desaparecer en la claridad del sol en 1801. Los astrónomos habían tratado

de estimar la trayectoria de Ceres, pero la forma en que tuvieron en cuenta el error de medición de sus telescopios era la equivocada, y el asteroide no reaparecía cerca de la ubicación que indicaban sus resultados. Gauss volvió a hacer sus cálculos utilizando el método de los mínimos cuadrados. Cuando los astrónomos orientaron sus telescopios hacia el punto que él había indicado, ¡encontraron a Ceres!

Científicos de distintas disciplinas se apresuraron a adoptar el método de los mínimos cuadrados. Más de dos siglos después, sigue siendo la forma estándar de evaluar los errores siempre que el objetivo sea la exactitud. La ponderación de los errores por su cuadrado es fundamental en estadística. En la inmensa mayoría de las aplicaciones de todas las disciplinas científicas, el ECM manda. Como pronto veremos, el método tiene sorprendentes implicaciones.

LAS ECUACIONES DE ERROR

El papel del sesgo y del ruido en el error se resume fácilmente en dos expresiones que llamaremos «las ecuaciones de error». La primera de estas descompone el error de una sola medición en los dos componentes que ya conocemos: el sesgo —el error medio— y un «error ruidoso» residual. El error ruidoso es positivo cuando el error es mayor que el sesgo, y negativo cuando es menor. El promedio de los errores ruidosos es cero. Nada nuevo en la primera ecuación de error.

$$\text{Error en una medición} = \text{sesgo} + \text{error ruidoso}$$

La segunda ecuación de error es una descomposición del ECM, la medida de error global que ahora hemos introducido. Utilizando un poco de álgebra elemental, se puede demostrar que el ECM es igual a la suma de los cuadrados del sesgo y del ruido.[2] (Recordemos que el ruido es la desviación estándar de las mediciones, que es idéntica a la desviación estándar de los errores ruidosos.) Por tanto:

$$\text{Error global (ECM)} = \text{sesgo}^2 + \text{ruido}^2$$

La forma de esta ecuación —una suma de dos cuadrados— puede recordarnos a otra archiconocida del colegio: el teorema de Pitágoras.

Como se recordará, en un triángulo rectángulo, la suma de los cuadrados de los dos lados más cortos (o catetos) es igual al cuadrado del más largo (o hipotenusa). Esto sugiere una sencilla visualización de la ecuación de error, en la que ECM, sesgo^2 y ruido^2 son las áreas de tres cuadrados adheridos a los lados de un triángulo rectángulo. La figura 7 muestra cómo el ECM (el área del cuadrado más oscuro) es igual a la suma de las áreas de los otros dos cuadrados. En la parte de la izquierda hay más ruido que sesgo, y, en la de la derecha, más sesgo que ruido. Sin embargo, el ECM es el mismo y la ecuación de error se mantiene en ambos casos.

Como indican tanto la expresión matemática como su representación visual, el sesgo y el ruido juegan papeles idénticos en la ecuación de error. Son independientes uno del otro y se ponderan por igual en la determinación del error global. (Nótese que utilizaremos una descomposición similar en una suma de cuadrados cuando analicemos los componentes del ruido en capítulos posteriores.)

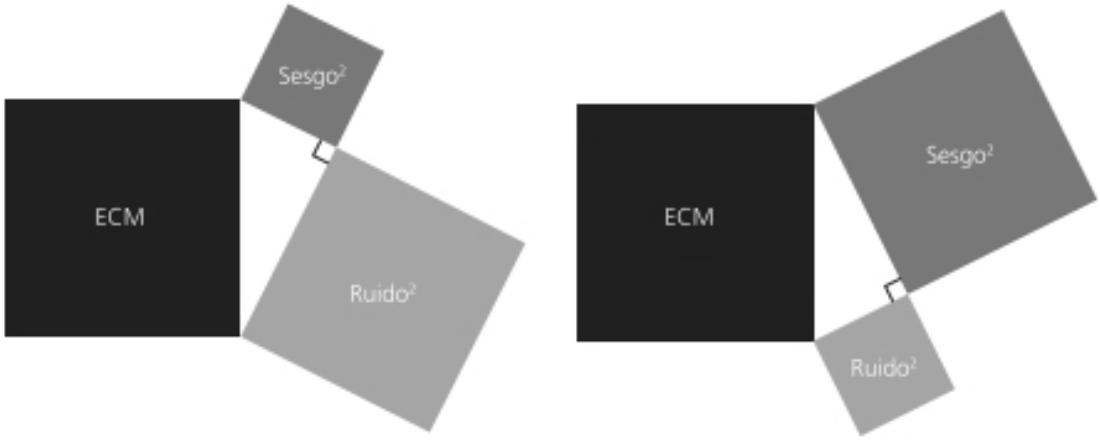

FIGURA 7. Dos descomposiciones del ECM

La ecuación de error proporciona una respuesta a la pregunta práctica que planteó Amy:¿cómo afectarán unas reducciones en igual cantidad del ruido o del sesgo al error global? La respuesta es sencilla: el sesgo y el ruido son intercambiables en la ecuación de error, y la disminución del error global será la misma, independientemente de cuál de los dos se reduzca. En la

figura 4, en la que el sesgo y el ruido resultan ser iguales (ambos son del 10 por ciento), sus contribuciones al error global son iguales.

La ecuación de error también proporciona un apoyo inequívoco al impulso inicial de Amy Simkin de intentar reducir el ruido. Siempre que se observe ruido, ¡hay que trabajar para reducirlo! La ecuación también demuestra que el jefe de Amy se equivocó cuando propuso que GoodSell esperara a medir el sesgo en sus pronósticos y solo entonces decidiría qué hacer. En términos de error global, el ruido y el sesgo son independientes: el beneficio de reducir el ruido es el mismo, con independencia de la cantidad de sesgo.

Esta noción es muy contraria a la intuición, pero es crucial. Para ilustrarlo, la figura 8 muestra el efecto de reducir el sesgo y el ruido en la misma cantidad. Para ayudar a apreciar lo que se ha conseguido en ambos paneles, la distribución original de los errores (de la figura 4) se representa con una línea discontinua.

En el panel A suponemos que el jefe de Amy decidió hacer las cosas a su manera: descubrió cuál era el sesgo y, de algún modo, consiguió reducirlo a la mitad (quizá proporcionando información a los analistas, demasiado optimistas). No se hizo nada con respecto al ruido. La mejora es visible: toda la distribución de los pronósticos se ha acercado al valor real.

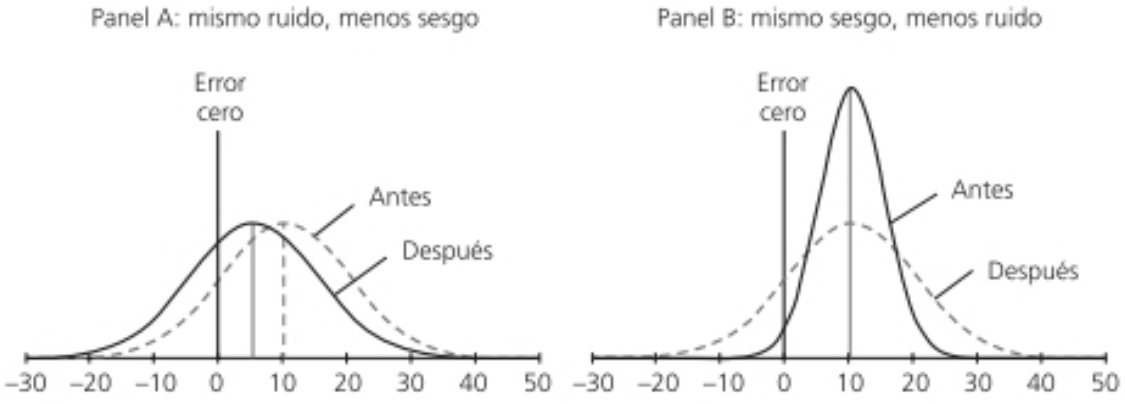

FIGURA 8. Distribución de errores con sesgo reducido a la mitad *versus* ruido reducido a la mitad

En el panel B, mostramos lo que habría ocurrido si Amy hubiera ganado en la discusión. El sesgo no cambia, pero el ruido se reduce a la mitad. La paradoja aquí es que la reducción del ruido parece haber empeorado las

cosas. Los pronósticos están ahora más concentrados (son menos ruidosos), pero no son más acertados (no están menos sesgados). Mientras que el 84 por ciento de los pronósticos se situaban a un lado del valor real, ahora casi todos (98 por ciento) se equivocan en la dirección de sobrepasar el valor real. La reducción del ruido parece haber hecho que los pronósticos estén equivocados con mayor precisión (¡difícilmente el tipo de mejora que esperaba Amy!).

Sin embargo, a pesar de las apariencias, el error global se ha reducido en el panel B tanto como en el A. La ilusión de deterioro en el panel B nace de una intuición errónea sobre el sesgo. La medición relevante del sesgo no es el desequilibrio entre los errores positivos y los negativos. Es el error medio, que en la figura 8 es la distancia entre el pico de la curva y el valor real. En el panel B, este error medio no ha cambiado con respecto a la situación original (sigue siendo alto, de un 10 por ciento, pero no es peor). Es cierto que la presencia del sesgo es ahora más llamativa, porque representa una proporción mayor del error global (el 80 por ciento en lugar del 50 por ciento), pero esto se debe a que el ruido se ha reducido. Por el contrario, en el panel A, el sesgo se ha reducido, pero no el ruido. El resultado neto es que el ECM es el mismo en ambos paneles: reducir el ruido o reducir el sesgo en la misma cantidad tiene el mismo efecto en el ECM.

Como este ejemplo ilustra, el ECM entra en conflicto con las intuiciones comunes en la puntuación de juicios predictivos. Para minimizar el ECM, hay que concentrarse en evitar los errores grandes. En la medición de una longitud, por ejemplo, el efecto de reducir un error de 11 cm a 10 cm es 21 veces mayor que el efecto de pasar de un error de 1 cm a un acierto completo. Desgraciadamente, las intuiciones de las personas a este respecto son casi la imagen especular de lo que deberían ser:[3] a la gente le interesa ante todo conseguir aciertos completos y es muy sensible a los pequeños errores, pero apenas le preocupa la diferencia entre dos grandes errores. Incluso si cree sinceramente que su objetivo es emitir juicios acertados, su reacción emocional a los resultados puede ser incompatible con el logro de la exactitud tal como la define la ciencia.

Por supuesto, la mejor solución en este caso sería reducir tanto el ruido como el sesgo. Dado que el sesgo y el ruido son independientes, no hay razón para elegir entre Amy Simkin y su jefe. En este sentido, si GoodSell

decide reducir el ruido, el hecho de que la reducción de ruido haga más visible el sesgo —de hecho, es imposible de pasar por alto— puede resultar una bendición. Lograr la reducción del ruido hará que la reducción del sesgo sea lo siguiente en la agenda de la compañía.

Es verdad que la reducción del ruido sería menos prioritaria si el sesgo fuese mucho mayor que el ruido, pero el ejemplo de GoodSell ofrece otra lección que merece la pena destacar. En este modelo simplificado hemos supuesto que el ruido y el sesgo son iguales. Dada la forma de la ecuación de error, sus contribuciones al error total también son iguales: el sesgo representa el 50 por ciento del error global, y lo mismo el ruido. Sin embargo, como hemos señalado, el 84 por ciento de los analistas se equivocan en la misma dirección. Hace falta un sesgo así de grande (¡seis de cada siete personas yerran en la misma dirección!) para que tenga tanto efecto como el ruido. Por tanto, no debería sorprendernos encontrar situaciones en las que hay más ruido que sesgo.

Hemos ilustrado la aplicación de la ecuación de error en un solo caso, una zona concreta del territorio de GoodSell. Por supuesto, siempre es deseable realizar una auditoría del ruido en varios casos a la vez. Nada cambia. La ecuación de error se aplica a los casos por separado, y se obtiene una ecuación global tomando los promedios del ECM, sesgo al cuadrado y ruido al cuadrado en los distintos casos. Habría sido mejor para Amy Simkin obtener múltiples pronósticos para varias zonas, ya sea de los mismos o de diferentes analistas. Promediar los resultados le daría una imagen más precisa del sesgo y del ruido en el sistema de predicción de GoodSell.

EL COSTE DEL RUIDO

La ecuación de error es el fundamento intelectual de este libro. Proporciona la justificación para reducir el ruido de sistema en los juicios predictivos, un objetivo que en principio es tan importante como la reducción del sesgo estadístico. (Debemos subrayar que «sesgo estadístico» no es sinónimo de «discriminación social»; es simplemente el error medio en un conjunto de juicios.)

La ecuación de error y las conclusiones que hemos extraído de ella dependen del uso del ECM como medida de error global. Esta regla es

adecuada para los juicios puramente predictivos, incluidos los pronósticos y las estimaciones, todos los cuales pretenden aproximarse a un valor real con la máxima exactitud (el menor sesgo) y precisión (el menor ruido).

Sin embargo, la ecuación de error no se aplica a los juicios evaluativos, porque el concepto de «error», que depende de la existencia de un valor real, es mucho más difícil de aplicar. Además, incluso si se pudieran especificar los errores, sus costes raramente serían simétricos, y sería improbable que fueran precisamente proporcionales a su cuadrado.

Para una empresa que fabrica ascensores, por ejemplo, las consecuencias de los errores en la estimación de la carga máxima de un ascensor son obviamente asimétricas: la subestimación es costosa, pero la sobrestimación podría ser catastrófica. El error al cuadrado es igualmente irrelevante para la decisión de cuándo salir de casa para tomar un tren. Para esta decisión, las consecuencias de llegar un minuto tarde o cinco minutos tarde son las mismas. Y, cuando la compañía de seguros del capítulo 2 fija el precio de las pólizas o estima el valor de las reclamaciones, los errores en ambas direcciones son costosos, pero no hay razón para suponer que sus costes sean equivalentes.

Estos ejemplos ponen de manifiesto la necesidad de especificar los papeles de los juicios predictivos y evaluativos en las decisiones. Una máxima ampliamente aceptada en la toma de decisiones es que no hay que mezclar los valores y los hechos. Una buena toma de decisiones debe basarse en juicios predictivos objetivos y precisos que no se vean en absoluto afectados por esperanzas y temores, o por preferencias y valores. Para la empresa de ascensores, el primer paso sería un cálculo neutral de la carga técnica máxima del ascensor para diferentes soluciones de la ingeniería. La seguridad es una consideración predominante solo en el segundo paso, cuando un juicio evaluativo determina la elección de un margen de seguridad aceptable para fijar la capacidad máxima. (Sin duda, esa elección también dependerá en gran medida de juicios objetivos que impliquen, por ejemplo, los costes y beneficios de ese margen de seguridad.) Del mismo modo, el primer paso para decidir cuándo salir camino de la estación debe ser una determinación objetiva de las probabilidades de los diferentes tiempos de viaje. Los costes respectivos de perder el tren y de perder el tiempo en la estación solo son relevantes a la hora de elegir el riesgo que se está dispuesto a aceptar.

La misma lógica es aplicable a decisiones con muchas más consecuencias. Un jefe militar debe sopesar muchas consideraciones a la hora de decidir si lanza una ofensiva, pero gran parte de la intelección en que confía es una cuestión de juicio predictivo. Un gobierno que responde a una crisis sanitaria, como una pandemia, debe sopesar los pros y los contras de las distintas opciones, pero ninguna evaluación es posible sin predicciones exactas sobre las probables consecuencias de cada opción (incluida la decisión de no hacer nada).

En todos estos ejemplos, las decisiones finales requieren juicios evaluativos. Los responsables de la toma de decisiones deben considerar varias opciones y aplicar sus valores para hacer la elección óptima. Sin embargo, las decisiones dependen de predicciones subyacentes, que deben ser valorativamente neutrales. Su objetivo es el acierto, es decir, dar lo más cerca posible de la diana, y el ECM es la medida de error adecuada. Los juicios predictivos mejorarán con procedimientos que reduzcan el ruido, siempre que no aumenten en exceso el sesgo.

SOBRE LA ECUACIÓN DE ERROR

- «Curiosamente, reducir el sesgo y el ruido en la misma cantidad tiene el mismo efecto en la exactitud.»
- «Siempre es útil reducir el ruido en el juicio predictivo, independientemente de lo que se sepa sobre el sesgo.»
- «Cuando unos juicios varían de 84 a 16 entre los que están por encima y por debajo del valor real, hay un gran sesgo (es cuando el sesgo y el ruido son iguales).»
- «Los juicios predictivos intervienen en todas las decisiones, y el acierto debe ser su único objetivo. Es preciso mantener separados valores y hechos.»

6
El análisis del ruido

En el capítulo anterior hablamos de la variabilidad en la medición o en el juicio de un solo caso. Cuando nos centramos en un solo caso, toda variabilidad del juicio es error, y los dos componentes del error son el sesgo y el ruido. Por supuesto, los sistemas de enjuiciamiento que estamos examinando, incluidos los que son propios de tribunales de justicia y compañías de seguros, están diseñados para tratar diferentes casos y distinguir entre ellos. Los jueces federales y los tasadores de reclamaciones no servirían de mucho si emitieran el mismo juicio para todos los casos que se les presentan. Gran parte de la variabilidad en los juicios de casos diferentes es deliberada.

Sin embargo, la variabilidad en los juicios de un mismo caso siempre será indeseada (es ruido del sistema). Como demostraremos, una auditoría del ruido en la que las mismas personas emiten juicios sobre varios casos permite un análisis más detallado del ruido del sistema.

UNA AUDITORÍA DEL RUIDO EN LAS SENTENCIAS

Para ilustrar el análisis del ruido con múltiples casos, recurriremos a una auditoría del ruido excepcionalmente detallada de sentencias de jueces federales. El análisis se publicó en 1981 como parte del movimiento por la reforma de las sentencias que describimos en el capítulo 1.[1] El estudio se centró en decisiones de sentencias, pero las lecciones que encierra son generales, y valen para otros juicios profesionales. El objetivo de la auditoría del ruido era ir más allá de las ochenta y ocho evidencias gráficas, pero anecdóticas, de ruido reunidas por el juez Frankel y otros y «determinar el alcance de la disparidad de las sentencias» de forma más sistemática.

Los autores del estudio elaboraron dieciséis casos hipotéticos en los que el acusado había sido declarado culpable y debía ser condenado. Las historias presentaban robos o casos de fraude, y diferían en otras seis

dimensiones, entre ellas si el acusado era autor material o cómplice del delito, si tenía antecedentes penales, si (para los casos de robo) había utilizado un arma, etcétera.

Los investigadores organizaron entrevistas cuidadosamente estructuradas con una muestra nacional de doscientos ocho jueces federales en activo. Durante noventa minutos se presentaron a los jueces los dieciséis casos y se les pidió que dictaran una sentencia.[2]

Para apreciar lo que se puede aprender de este estudio, será útil un ejercicio de visualización. Imaginemos una gran tabla con 16 columnas para los delitos, etiquetadas de la A a la P, y 208 filas para los jueces, numeradas de 1 a 208. Cada celda, de la A1 a la P208, muestra la pena de prisión dictada para un caso concreto por un juez determinado. La figura 9 ilustra el aspecto de esta tabla de 3328 celdas. Para estudiar el ruido, nos centraremos en las 16 columnas, cada una de las cuales es una auditoría del ruido independiente.

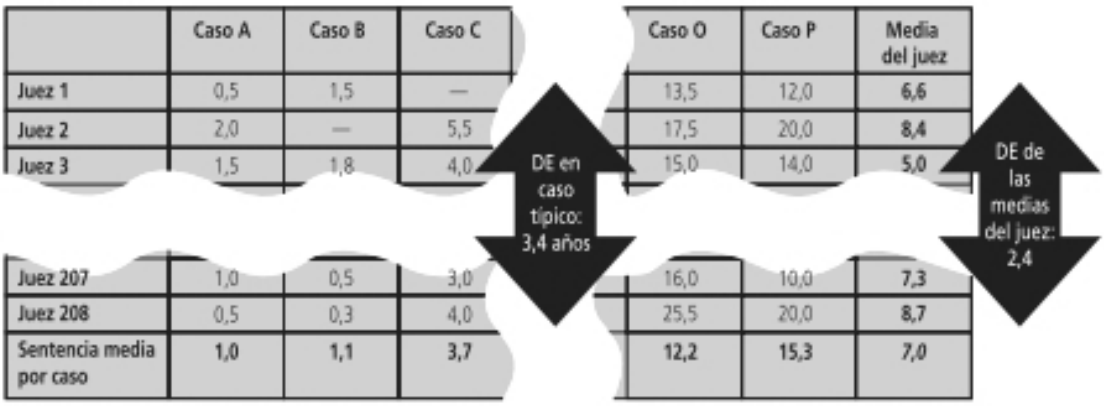

FIGURA 9. Una representación del estudio de las sentencias

SENTENCIAS MEDIAS

No existe una forma objetiva de determinar cuál es el «valor real» de una sentencia para un caso particular. En lo que sigue, tratamos el promedio de las 208 sentencias para cada caso (sentencia media) como si fuera la sentencia «justa» para ese caso. Como señalamos en el capítulo 1, la Comisión de Sentencias de Estados Unidos hizo la misma suposición cuando utilizó la práctica promedial en casos anteriores como base para

establecer las directrices que debían seguir las sentencias. Esta sentencia «justa» supone un sesgo cero en el juicio medio de cada caso.

Somos plenamente conscientes de que, en realidad, esta suposición es errónea: es muy probable que el juicio promedio de algunos casos esté sesgado en relación con el juicio promedio de otros casos muy similares, por ejemplo, debido a discriminación racial. La varianza de los sesgos entre los casos —unos positivos, otros negativos— es una importante fuente de error e injusticia. Es frecuente que a esta varianza se la llame confusamente «sesgo».[3] Nuestro análisis en este capítulo —y en este libro— se centra en el ruido, que es una fuente de error distinta. El juez Frankel hizo hincapié en la injusticia del ruido, pero también llamó la atención sobre el sesgo (incluida la discriminación racial). Del mismo modo, nuestro foco puesto en el ruido no debe interpretarse como una subestimación de la importancia de medir y combatir sesgos compartidos.

Por comodidad, en la fila inferior de la tabla se indica la sentencia media de cada caso. Los casos están ordenados por orden creciente de severidad: la sentencia media en el caso A es de 1 año; en el caso P es de 15,3 años. La media de las penas de prisión para los 16 casos es de 7 años.[4]

Imaginemos ahora un mundo perfecto en el que todos los jueces son instrumentos precisos de medición de la justicia y las sentencias están libres de ruido. ¿Cómo sería la figura 9 en un mundo así? Evidentemente, todas las celdas de la columna del caso A serían idénticas, porque todos los jueces darían al acusado del caso A la misma sentencia de exactamente un año. Igual sucedería en todas las demás columnas. Los números de cada fila, por supuesto, seguirían variando, porque los casos son diferentes, pero cada fila sería idéntica a la de arriba y a la de abajo. Las diferencias entre los casos serían la única fuente de variabilidad en la tabla.

Desafortunadamente, el mundo de la justicia federal no es perfecto. Los jueces no son idénticos, y la variabilidad dentro de las columnas es grande, lo que indica que hay ruido en los juicios de cada caso. La variabilidad de las sentencias es mayor de lo que debería ser, y el objetivo del estudio es analizarla.

Partamos de la imagen del mundo perfecto que hemos descrito antes, un mundo en el que todos los casos reciben la misma pena por parte de todos los jueces. Cada columna es una serie de 208 números idénticos. Ahora, añadamos ruido bajando por cada columna y cambiando algunos números aquí y allá, a veces añadiendo tiempo de prisión a la sentencia media, a veces restándolo. Como los cambios que se realizan no son todos iguales, crean variabilidad dentro de la columna. Esta variabilidad es ruido.

El resultado esencial de este estudio es la gran cantidad de ruido que se observa «en los juicios de cada caso». La medida del ruido en cada caso es la desviación estándar de las penas de prisión asignadas a ese caso. En el caso medio, la sentencia media fue de 7 años, y la desviación estándar en torno a esa media fue de 3,4 años.[5]

Aunque es posible que el lector esté familiarizado con el término «desviación estándar», puede encontrar útil una descripción concreta. Imagine que elige al azar a dos jueces y calcula la diferencia entre sus juicios de un caso. Ahora repita la operación para todos los pares de jueces y todos los casos, y promedie los resultados. Esta medida, la «diferencia media absoluta», debería darle una idea de la lotería a la que se enfrenta el acusado en un tribunal federal. Suponiendo que las sentencias se distribuyan normalmente, es 1,128 veces la desviación estándar, lo que implica que la diferencia media entre dos sentencias elegidas al azar del mismo caso será de 3,8 años. En el capítulo 3 hablamos de la lotería a la que se enfrenta el cliente que necesita una suscripción especializada de una compañía de seguros. La lotería penal del acusado tiene, por no decir otra cosa, más consecuencias.

Una diferencia media absoluta de 3,8 años entre los jueces cuando la condena media es de 7 años es un resultado preocupante y, en nuestra opinión, inaceptable. Sin embargo, hay buenas razones para sospechar que hay aún más ruido en la administración real de justicia. En primer lugar, los participantes en la auditoría del ruido se enfrentaron a casos artificiales, que eran inusualmente fáciles de comparar y se presentaban en sucesión inmediata. La vida real no ofrece tanto apoyo para mantener la coherencia. En segundo lugar, los jueces de un tribunal tienen mucha más información que la que tenían aquí. La nueva información, a menos que sea decisiva, proporciona más ocasiones para que los jueces difieran entre sí. Por estas

razones, sospechamos que la cantidad de ruido a la que se enfrentan los acusados en las salas reales es aún mayor que la que vemos aquí.

ALGUNOS JUECES SON SEVEROS: RUIDO DE NIVEL

En el siguiente paso del análisis, los autores desglosaron el ruido en componentes separados. La primera interpretación del ruido que probablemente se les ocurrió —como al juez Frankel— es que el ruido se debe a variación entre los jueces en su disposición a dictar sentencias severas. Como cualquier abogado defensor dirá, los jueces tienen una reputación: unos la de ser duros «jueces de la horca», más severos que el juez medio, y otros la de ser «jueces con corazón», más indulgentes que el juez medio. Nos referimos a estas desviaciones como «errores de nivel». (Nuevamente se define aquí el «error» como una desviación de la media; un error puede de hecho corregir una injusticia si el juez medio está equivocado.)

La variabilidad en los errores de nivel se encontrará en cualquier otra tarea que suponga juzgar. Ejemplos de ello son las evaluaciones de rendimiento, en las que algunos supervisores son más generosos que otros; las predicciones de cuota de mercado, en las que algunos analistas son más optimistas que otros, o las recomendaciones de cirugía de la espalda, en las que algunos ortopedistas son más agresivos que otros.

Cada fila de la figura 9 muestra las sentencias dictadas por un juez. La pena media fijada por cada juez, que se muestra en la columna de la derecha de la tabla, es una medida del nivel de severidad del juez. Se observa que los jueces varían mucho en esta dimensión. La desviación estándar de los valores de la última columna de la derecha fue de 2,4 años. Esta variabilidad no tiene nada que ver con la justicia. En cambio, como cabe sospechar, las diferencias en la media de las sentencias reflejan la variación entre los jueces en otras características: sus antecedentes, las experiencias de su vida, sus opiniones políticas, sesgos, etcétera. Los investigadores examinaron las actitudes de los jueces con respecto a las sentencias en general; por ejemplo, si creen que el objetivo principal de las sentencias es la incapacitación (apartar al delincuente de la sociedad), la rehabilitación o la disuasión. Descubrieron que los jueces que piensan que el objetivo principal es la rehabilitación tienden a asignar penas de prisión más cortas y

más tiempo de supervisión que los jueces que apuntan a la disuasión o a la incapacitación. Por otra parte, los jueces ubicados en el sur de Estados Unidos asignaron sentencias significativamente más largas que sus homólogos en otras partes del país. No es sorprendente que la ideología conservadora también esté relacionada con la severidad de las sentencias.

La conclusión general es que el nivel medio de condena funciona como un rasgo de personalidad. Se podría utilizar este estudio para clasificar a los jueces en una escala que vaya de muy duro a muy indulgente, del mismo modo que un test de personalidad puede medir el grado de extraversión o de simpatía. Al igual que otros rasgos, cabría esperar que la severidad de la sentencia estuviera correlacionada con factores genéticos, con experiencias vitales y con otros aspectos de la personalidad. Nada de esto tiene que ver con el caso o con el acusado. Utilizamos el término «ruido de nivel» para la variabilidad de los juicios medios de los jueces, que es idéntica a la variabilidad de los errores de nivel.

LOS JUECES DIFIEREN: RUIDO DE PATRÓN

Como muestran las flechas negras de la figura 9, el ruido de nivel es de 2,4 años, y el del sistema, de 3,4 años. Esta diferencia indica que en el ruido del sistema hay algo más que las diferencias en la severidad media de los jueces individuales. A este otro componente del ruido lo llamamos «ruido de patrón».

Para entender el ruido de patrón, consideremos de nuevo la figura 9 y centrémonos en una celda elegida al azar, por ejemplo, la celda C3. La sentencia justa en el caso C se muestra en la parte inferior de la columna; como puede verse, es de 3,7 años. Ahora, fijémonos en la columna de la derecha para encontrar la sentencia media dada por el juez 3 en todos los casos. Es de 5 años, solo 2 años menos que la media general. Si la variación en la severidad de los jueces fuese la única fuente de ruido en la columna 3, se predeciría que la sentencia en la celda C3 es de 3,7 – 2 = 1,7 años. Sin embargo, el dato real en la celda C3 es de 4 años, lo que indica que el juez 3 fue especialmente duro al sentenciar ese caso.

La misma simple y aditiva lógica permitiría predecir todas las sentencias de todas las columnas de la tabla, pero de hecho se hallarán desviaciones del modelo simple en la mayoría de las celdas.[6] Si se observa una fila, se

encontrará que los jueces no son igualmente severos en sus sentencias de todos los casos: son más duros que su media personal en algunos y más indulgentes en otros. Llamamos a estas desviaciones residuales «errores de patrón». Si se anotan estos errores de patrón en cada celda de la tabla, se verá que suman cero para cada juez (fila) y que también suman cero para cada caso (columna). Sin embargo, los errores de patrón no se eliminan en su contribución al ruido, porque los valores de todas las celdas se hallan elevados al cuadrado para el cálculo del ruido.

Hay un modo más fácil de confirmar que el simple modelo aditivo de las sentencias no se sostiene. En la tabla puede verse que las sentencias medias en la parte inferior de cada columna aumentan de forma constante de izquierda a derecha, pero no ocurre lo mismo dentro de las filas. El juez 208, por ejemplo, fijó una pena mucho más alta para el acusado del caso O que para el del caso P. Si los jueces individuales clasificaran los casos por la severidad de la pena apropiada, sus clasificaciones no serían las mismas.

Utilizamos el término «ruido de patrón» para la variabilidad que acabamos de identificar, porque esa variabilidad refleja un patrón complejo en las actitudes de los jueces hacia casos particulares. Un juez, por ejemplo, puede ser más duro que la media en general, pero algo más indulgente con los delincuentes de cuello blanco. Otro puede estar inclinado a castigar con ligereza, pero con mayor severidad cuando el delincuente reincide. Y un tercero puede estar cerca de la severidad media, pero ser comprensivo cuando el delincuente es un mero cómplice y duro cuando la víctima es una persona de edad avanzada. (Utilizamos el término «ruido de patrón» en aras de la legibilidad. El término estadístico más propio para el ruido de patrón es «interacción juez × caso». Pedimos disculpas a las personas con formación estadística por imponerles la carga de la traducción.)

En el contexto de la justicia penal, algunas de las reacciones idiosincrásicas a los casos pueden corresponder a la filosofía personal del juez a la hora de dictar sentencias. Otras respuestas pueden provenir de asociaciones de las que el juez apenas es consciente, como un acusado que le recuerda a un delincuente especialmente odioso o una acusada que se parece a su hija. Sea cual fuere su origen, estas pautas no son mera casualidad: cabría esperar que se repitieran si el juez volviera a ver el mismo caso. Sin embargo, como el ruido de patrón es, en la práctica, difícil de predecir, añade incertidumbre a la ya impredecible lotería de las

sentencias. Como señalaban los autores del estudio, «las diferencias de patrón entre los jueces en cuanto a la influencia de las características del delito/delincuente» son «una forma adicional de disparidad en las sentencias».[7]

Puede que el lector haya notado que la descomposición del ruido del sistema en ruido de nivel y ruido de patrón sigue la misma lógica que la ecuación de error del capítulo anterior, que descomponía el error en sesgo y ruido. Esta vez, la ecuación puede escribirse como sigue:

$$\text{Ruido del sistema}^2 = \text{ruido de nivel}^2 + \text{ruido de patrón}^2$$

Esta expresión puede representarse visualmente de la misma manera que la ecuación de error original (figura 10). Hemos representado los dos lados del triángulo como iguales. Esto se debe a que, en el estudio de las sentencias, el ruido de patrón y el ruido de nivel contribuyen casi por igual al ruido del sistema.[8]

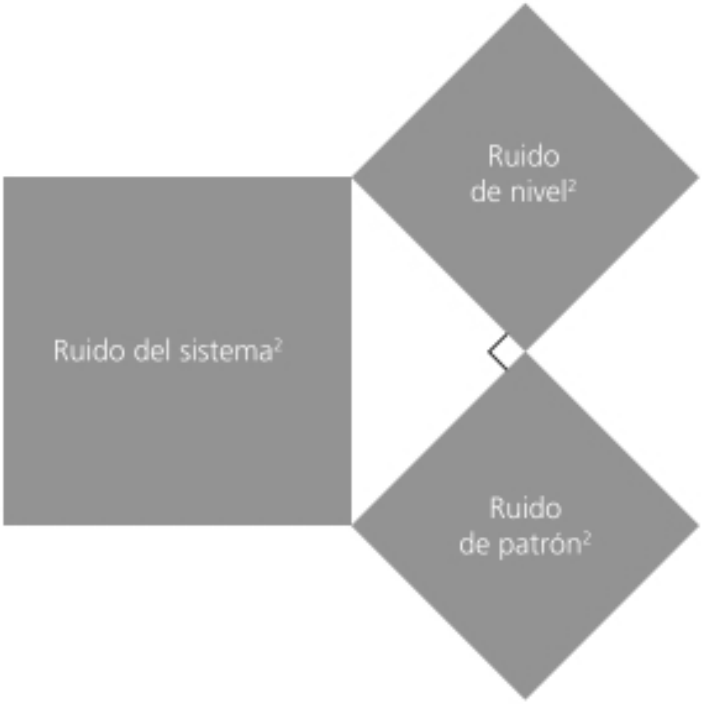

FIGURA 10. Descomposición del ruido del sistema

El ruido de patrón es omnipresente. Los médicos deciden si admiten a ciertas personas para su hospitalización, las empresas deciden a quién

contratar, los abogados deciden qué casos llevar o los ejecutivos de Hollywood deciden qué programas de televisión producir. En todos estos casos, hay ruido de patrón, con diferentes juzgadores que hacen distintas clasificaciones de los casos.

LOS COMPONENTES DEL RUIDO

Nuestro tratamiento del ruido de patrón pasó por encima de una complejidad importante: la posible contribución del error aleatorio.

Recuerde el lector el ejercicio del cronómetro. Cuando trató de medir diez segundos repetidamente, los resultados variaron de una vuelta a otra; mostraron una variabilidad dentro de su persona. Por la misma razón, los jueces no habrían fijado exactamente las mismas sentencias a los dieciséis casos si se les hubiera pedido que los juzgaran de nuevo. Como veremos, no habrían dictado las mismas sentencias si el estudio original se hubiera realizado otro día de la misma semana. Si un juez está de buen humor porque le ha pasado algo bueno a su hija, o porque su equipo deportivo favorito ganó el día anterior, o porque hace un día soleado, su juicio puede ser más indulgente de lo que sería en otras circunstancias. Esta variabilidad intrapersonal es conceptualmente distinta de las diferencias estables entre personas que acabamos de analizar, pero es difícil distinguir estas fuentes de variabilidad. Nuestro término para la variabilidad que se debe a los efectos transitorios es «ruido de ocasión».

De este estudio ignoramos, en efecto, el ruido de ocasión y optamos por interpretar los patrones idiosincrásicos de los jueces en la auditoría del ruido como indicadores de actitudes estables. Esta suposición es con toda certeza optimista, pero hay razones independientes para creer que el ruido de ocasión no desempeñó un papel importante en este estudio. Los jueces altamente experimentados que participaron en él seguro que traían consigo algunas ideas fijas sobre la importancia de las distintas características de los delitos y de los acusados. En el próximo capítulo trataremos con más detalle el ruido de ocasión y mostraremos cómo se puede separar del componente estable de ruido de patrón.

En resumen, hemos hablado de varios tipos de ruido. El «ruido del sistema» es una variabilidad no deseada en los juicios de un mismo caso por parte de varios individuos. Hemos identificado sus dos componentes

principales, que pueden separarse cuando los mismos individuos evalúan varios casos:

• El «ruido de nivel» es variabilidad en el nivel medio de los juicios de los distintos jueces.
• El «ruido de patrón» es variabilidad de las respuestas de los jueces a casos concretos.

En el presente estudio, las cantidades de ruido de nivel y de ruido de patrón eran casi iguales. Sin embargo, el componente que identificamos como ruido de patrón contiene sin duda algo de «ruido de ocasión», que puede tratarse como un error aleatorio.

Hemos utilizado una auditoría del ruido en el sistema judicial como ejemplo, pero el mismo análisis puede aplicarse a cualquier auditoría del ruido, en empresas, medicina, educación, gobierno o cualquier otra área. El ruido de nivel, el ruido de patrón estable y el ruido de ocasión contribuyen al ruido del sistema, y nos encontraremos con ellos repetidamente conforme avancemos.

SOBRE ANALIZAR EL RUIDO

- «El ruido de nivel se produce cuando los jueces muestran diferentes niveles de gravedad. El ruido de patrón se produce cuando no están de acuerdo entre ellos sobre qué acusados merecen un tratamiento más severo o más indulgente. Y parte del ruido de patrón es ruido de ocasión (cuando los jueces no están de acuerdo consigo mismos).»
- «En un mundo perfecto, los acusados se enfrentarían a la justicia; en nuestro mundo, se enfrentan a un sistema ruidoso.»

7
Ruido de ocasión

Un jugador profesional de baloncesto se prepara para lanzar un tiro libre. Se sitúa en la línea, se concentra y lanza el balón. Esta es una secuencia precisa de movimientos que ha practicado innumerables veces. ¿Encestará?

No lo sabemos, y él tampoco. En la NBA, los jugadores suelen encestar tres cuartas partes de sus intentos. Obviamente, unos jugadores son mejores que otros, pero ningún jugador anota el cien por cien de las veces. Los mejores de todos los tiempos lo han hecho en poco más del 90 por ciento de sus tiros libres. (En el momento de escribir estas líneas, son Stephen «Steph» Curry, Steve Nash y Mark Price.)[1] Los peores de todos los tiempos rondan el 50 por ciento. (El gran Shaquille O'Neal, por ejemplo, encestó siquiera el 53 por ciento de sus lanzamientos.)[2] Aunque la cesta está siempre a tres metros de altura y a cuatro metros de distancia exactos, y el balón siempre pesa 623,69 gramos, la capacidad de repetir la secuencia precisa de gestos necesaria para encestar no resulta fácil. No solo se espera que haya variabilidad entre los jugadores, sino también en cada uno. El tiro libre es una forma de lotería, con muchas más posibilidades de éxito si el lanzador es Curry que si es O'Neal, pero no deja de ser una lotería.

¿De dónde viene esta variabilidad? Sabemos que son incontables los factores que pueden influir en el jugador en la línea de tiro: el cansancio de un partido largo, la presión mental de un marcador ajustado, las aclamaciones de la afición local o los abucheos del equipo contrario. Si alguien como Curry o Nash falla, invocaremos una de estas explicaciones, pero, en realidad, es poco probable que sepamos el papel exacto de estos factores. La variabilidad en el rendimiento de un lanzador es una forma de ruido.

LA SEGUNDA LOTERÍA

La variabilidad en los tiros libres, o en otros procesos físicos, no es una sorpresa. Estamos acostumbrados a la variabilidad en nuestro cuerpo:

nuestras pulsaciones, nuestra presión arterial, nuestros reflejos, el tono de nuestra voz y el temblor de nuestras manos son diferentes en distintos momentos. Y, por mucho que nos esforcemos en escribir la misma firma, siempre será un poco diferente en cada cheque.

Menos fácil es observar la variabilidad de nuestras mentes. Por supuesto, todos hemos tenido la experiencia de cambiar de opinión, incluso sin nueva información. La película que una noche nos hizo reír a carcajadas nos parece hoy mediocre y la olvidaremos. La persona a la que ayer juzgamos con severidad parece merecer ahora nuestra indulgencia. Un razonamiento que no nos había gustado, o que no habíamos entendido, ahora lo entendemos y nos parece importante. Sin embargo, como estos ejemplos indican, solemos asociar estos cambios a cuestiones relativamente menores y en gran medida subjetivas.

En realidad, nuestras opiniones cambian sin razón aparente. Es algo que se da incluso en los juicios cuidadosos y atentos de expertos profesionales. Por ejemplo, es habitual obtener diagnósticos significativamente diferentes de los mismos médicos cuando se les presenta dos veces el mismo caso (véase el capítulo 22). Cuando los expertos en vinos de un importante concurso de Estados Unidos cataron dos veces los mismos vinos, solo puntuaron de forma idéntica el 18 por ciento de los vinos (en general, los peores).[3] Un experto forense puede llegar a conclusiones diferentes cuando examina dos veces las mismas huellas dactilares con solo unas semanas de diferencia (véase el capítulo 20). Consultores experimentados de *software* pueden ofrecer estimaciones muy diferentes del tiempo de realización de la misma tarea en dos ocasiones.[4] En resumen: del mismo modo que un jugador de baloncesto nunca lanza el balón dos veces exactamente de la misma manera, no siempre producimos juicios idénticos cuando nos enfrentamos a los mismos hechos en dos ocasiones.

Hemos descrito el proceso que elige un asegurador, un juez o un médico como una lotería que crea ruido del sistema. El ruido ocasional es el producto de una segunda lotería. Esta lotería escoge el momento en que el profesional se dispone a emitir un juicio, su estado de ánimo, la secuencia de casos que tiene en la memoria y un sinfín de otros aspectos de la ocasión. Esta segunda lotería suele ser mucho más abstracta que la primera. Podemos ver cómo la primera lotería podría haber seleccionado a un agente de seguros diferente, por ejemplo, pero las alternativas a las respuestas

reales del agente seleccionado son abstractamente contrafácticas. Solo sabemos que el juicio que se produjo fue escogido entre una nube de posibilidades. El ruido de ocasión es la variabilidad entre estas posibilidades no vistas.

MEDIR EL RUIDO DE OCASIÓN

Medir el ruido de ocasión no es fácil, por la misma razón que su existencia, una vez establecida, a menudo nos sorprende. Cuando las personas se forman una opinión profesional cuidadosamente considerada, la asocian a las razones que justifican su punto de vista. Si se las presiona para que expliquen su criterio, suelen defenderlo con razonamientos que les resultan convincentes. Y, si se les presenta el mismo asunto por segunda vez y lo reconocen, reproducirán la respuesta anterior para minimizar el esfuerzo y mantener la coherencia. Consideremos este ejemplo de la profesión docente: si un profesor pone la nota máxima a la redacción de un alumno y luego vuelve a leer la misma redacción una semana más tarde, es poco probable que, después de ver la nota original, le ponga una nota muy distinta.

Esta es la razón de que sea tan difícil hacer mediciones directas del ruido de ocasión cuando los casos son de fácil recuerdo. Si, por ejemplo, enseñamos a un agente de seguros, o a un juez penal, un caso que ha resuelto antes, lo más probable es que reconozca los hechos del caso y que se limite a repetir su juicio anterior. Una revisión de la investigación sobre la variabilidad del juicio profesional (técnicamente conocida como «fiabilidad test-retest» o, para abreviar, «fiabilidad») incluyó muchos estudios en los que los expertos emitieron el mismo juicio dos veces en la misma sesión. No es de extrañar que tendieran a estar de acuerdo consigo mismos.[5]

Los experimentos que mencionamos antes evitaron este problema utilizando estímulos que los expertos no reconocerían. Los jueces de los vinos participaron en una cata a ciegas. A los dactiloscopistas se les mostraron pares de huellas que ya habían visto y a los expertos en *software* se les preguntó por tareas en las que ya habían trabajado, pero semanas o meses después, y sin que se les dijera que se trataba de casos que ya habían examinado.

Hay otra forma menos directa de confirmar la existencia del ruido de ocasión: utilizar *big data* y métodos econométricos. Cuando se dispone de una amplia muestra de decisiones profesionales pasadas, los analistas a veces pueden comprobar si estas decisiones se vieron influidas por factores irrelevantes específicos de la ocasión, como la hora del día o la temperatura exterior. Los efectos estadísticamente significativos de tales factores irrelevantes en los juicios son una prueba del ruido de ocasión. Siendo realistas, no hay esperanza de descubrir todas las extrañas fuentes de ruido de ocasión, pero las que se pueden encontrar ilustran la gran variedad de estas fuentes. Para controlar el ruido de ocasión, debemos tratar de entender los mecanismos que lo producen.

UNO ES MULTITUD

Pensemos en esta pregunta: ¿qué porcentaje de los aeropuertos del mundo se encuentra en Estados Unidos? A quien la piense, puede que le venga a la mente una respuesta, pero no le vendrá de la misma manera que cuando recuerda su edad o su número de teléfono. Es consciente de que la cifra que acaba de dar es una estimación. No es un número aleatorio (el 1 por ciento o el 99 por ciento serían claramente respuestas erróneas). Sin embargo, la cifra que se le habrá ocurrido es solo una de un rango de posibilidades que no excluiría. Si alguien sumara o restara un punto porcentual a su respuesta, probablemente no le parecería que la estimación resultante fuera mucho menos plausible que la suya. (La respuesta correcta, en caso de que se lo pregunte, es el 32 por ciento.)[6]

Dos investigadores, Edward Vul y Harold Pashler, tuvieron la idea de pedir a una muestra de personas que respondieran a esta pregunta (y a muchas otras similares) no una, sino dos veces.[7] No se les dijo la primera vez que tendrían que volver a hacer una estimación. La hipótesis de Vul y Pashler era que la media de las dos respuestas sería más exacta que cualquiera de ellas por separado.

Los datos les dieron la razón. En general, la primera estimación se acercaba más a la verdad que la segunda, pero la mejor estimación se obtuvo al promediar las dos estimaciones.

Vul y Pashler se inspiraron en el conocido fenómeno del «efecto de la sabiduría de la multitud»: promediar los juicios independientes de distintas

personas suele mejorar la exactitud. En 1907, Francis Galton, primo de Darwin y famoso erudito, pidió a setecientos ochenta y siete aldeanos de una feria rural que estimaran el peso de un buey premiado. Ninguno de los aldeanos adivinó el peso real del buey, que era de 543 kilos, pero la media de sus estimaciones fue de 544 kilos, con solo un kilo de diferencia, y la mediana (547,5) también estuvo muy cerca. Los aldeanos eran una «multitud sabia» en el sentido de que, aunque sus cálculos individuales eran bastante ruidosos, no estaban sesgados. La demostración de Galton fue una sorpresa incluso para él: tenía poco respeto por el juicio de la gente corriente y, a su pesar, declaró que sus resultados eran «más loables en su exactitud de lo que cabía esperar de un juicio democrático».

Se han encontrado resultados similares en cientos de situaciones. Por supuesto, si las preguntas son tan difíciles que solo los expertos pueden acercarse a la respuesta, las multitudes no serán necesariamente muy exactas. Sin embargo, cuando, por ejemplo, se pide a la gente que adivine el número de gominolas que hay en un frasco transparente, que prediga la temperatura de su ciudad dentro de una semana o que calcule los beneficios por acción de distintas acciones, es probable que la respuesta media de un gran número de personas se aproxime a la verdad.[8] La razón es de estadística básica: al promediar varios juicios (o mediciones) independientes, se obtiene un nuevo juicio, que es menos ruidoso, aunque no menos sesgado, que los juicios individuales.[9]

Vul y Pashler querían averiguar si el mismo efecto se extiende al ruido de ocasión: ¿se puede estar más cerca de la verdad combinando dos estimaciones de la misma persona, igual que cuando se combinan las estimaciones de diferentes personas? Descubrieron que la respuesta era afirmativa. Vul y Pashler dieron a este hallazgo un nombre evocador: «la multitud interior».

Promediar dos estimaciones de la misma persona no mejora los juicios tanto como buscar una segunda opinión independiente. Como dicen Vul y Pashler, «se puede ganar, aproximadamente, 1/10 más de lo que se gana haciéndose uno la misma pregunta dos veces si se obtiene una segunda opinión de otra persona».[10] No es una gran mejora, pero puede lograr que el efecto sea mucho mayor si se espera a hacer una segunda estimación. Cuando Vul y Pashler dejaron pasar tres semanas antes de volver a realizar la misma pregunta a sus sujetos, el beneficio aumentó en un tercio del valor

de una segunda opinión. No está mal para una técnica que no requiere ninguna información añadida ni ayuda externa. Y este resultado justifica sin duda el viejo consejo a los responsables de la toma de decisiones: «Consúltalo con la almohada y vuelve a pensarlo por la mañana».

Trabajando independientemente de Vul y Pashler, pero más o menos al mismo tiempo, dos investigadores alemanes, Stefan M. Herzog y Ralph Hertwig, propusieron una puesta en práctica diferente del mismo principio. [11] En lugar de limitarse a pedir a sus sujetos que realizaran una segunda estimación, los animaron a generar una estimación que —aunque siguiera siendo plausible— fuera lo más diferente posible de la primera. Esta petición requería que los sujetos pensaran activamente en una información que no habían considerado la primera vez. Las instrucciones para los participantes eran las siguientes:

> En primer lugar, suponga que su primera estimación es incorrecta. En segundo lugar, piense en algunas razones por las que podría serlo. ¿Qué suposiciones y consideraciones podrían haber sido erróneas? En tercer lugar, ¿qué implican estas nuevas consideraciones? ¿La primera estimación era demasiado alta o demasiado baja? En cuarto lugar, haga desde esta nueva perspectiva una segunda estimación alternativa.

Al igual que Vul y Pashler, Herzog y Hertwig calcularon la media de las dos estimaciones hechas así. Su técnica, que denominaron «*bootstrapping* dialéctico», produjo mayores mejoras en la exactitud que una simple solicitud de una segunda estimación inmediatamente después de la primera. Como los participantes se obligaron a considerar la cuestión bajo una nueva luz, tomaron como muestra otra versión más diferente de sí mismos, dos «miembros» de la «multitud interior» que estuvieran más alejados. El resultado fue que su promedio produjo una estimación más próxima a la verdad. La ganancia de acierto con dos estimaciones «dialécticas» inmediatamente consecutivas fue casi de la mitad del valor de una segunda opinión.

El resultado para los responsables de la toma de decisiones, tal y como lo resumen Herzog y Hertwig, es una simple elección entre procedimientos: si puede reunir opiniones independientes de otros, hágalo; es muy probable que esta auténtica sabiduría de la multitud mejore su juicio. Si no puede, haga su mismo juicio una segunda vez para crear una «multitud interior». Puede hacerlo cuando haya pasado algún tiempo —y se haya distanciado de

su primera opinión— o intentando argumentar activamente contra sí mismo y buscar otra perspectiva para el problema. Por último, lo más indicado es que haga una media independientemente del tipo de multitud (a menos que tenga poderosas razones para dar más peso a una de las estimaciones).

Más allá de los consejos prácticos, esta línea de investigación confirma una idea esencial sobre el juicio. Como dicen Vul y Pashler, «las respuestas de un sujeto se muestrean a partir de una distribución interna de probabilidad, en lugar de seleccionarse de forma determinista a partir de todos los conocimientos que tiene un sujeto».[12] Esta observación se hace eco de la experiencia que tuvo el lector al responder a la pregunta sobre los aeropuertos de Estados Unidos: su primera respuesta no recogía todos sus conocimientos, ni siquiera los mejores. La respuesta era solo un punto en la nube de posibles respuestas que su mente pudo haber generado. La variabilidad que observamos en los juicios del mismo problema por la misma persona no es una casualidad observada en unos pocos problemas muy especializados: el ruido de ocasión afecta siempre a todos nuestros juicios.

LAS FUENTES DEL RUIDO DE OCASIÓN

Hay al menos una fuente de ruido de ocasión que todos hemos notado: el estado de ánimo. Todos hemos experimentado cómo nuestros propios juicios pueden depender de cómo nos sintamos (y, desde luego, somos conscientes de que los juicios de los demás también varían según su estado de ánimo).

El efecto de los estados de ánimo en el juicio ha sido objeto de una enorme cantidad de investigaciones psicológicas. Es muy fácil hacer que las personas estén temporalmente contentas o tristes y medir la variabilidad de sus juicios y decisiones después de haberles inducido esos estados de ánimo. Para ello, los investigadores utilizan varias técnicas. Por ejemplo, a veces se pide a los participantes que escriban un párrafo recordando un acontecimiento feliz o uno triste. A veces, solo ven una secuencia de vídeo extraída de una película divertida o de otra melodramática.

Varios psicólogos llevan décadas investigando los efectos de la manipulación del estado de ánimo. Quizá el más prolífico sea el psicólogo

australiano Joseph P. Forgas.[13] Ha publicado alrededor de un centenar de artículos científicos sobre el tema del estado de ánimo.

Algunas de las investigaciones de Forgas confirman lo que todo el mundo piensa: las personas que están de buen humor suelen ser más positivas. Les resulta más fácil recordar los momentos felices que los tristes, aceptan más a otras personas, son más generosas y serviciales, etcétera. El estado de ánimo negativo tiene los efectos contrarios. Como escribe Forgas, «la misma sonrisa que es vista como amistosa por una persona de buen humor puede ser juzgada sospechosa cuando el observador está de mal humor; hablar del tiempo puede ser visto como algo natural cuando la persona está de buen humor, pero aburrido cuando está de mal humor».[14]

En otras palabras, el estado anímico tiene una influencia medible en lo que uno piensa: lo que nota en su entorno, lo que recupera de su memoria y cómo da sentido a estas señales. Pero el estado anímico tiene otro efecto más sorprendente: también cambia la «forma» de pensar. Y, en este caso, los efectos no son los que cabe imaginar. Estar de buen humor tiene sus pros y sus contras, y el mal humor tiene su lado bueno. Los costes y beneficios de los distintos estados de ánimo son específicos de cada situación.

En un contexto de negociación, por ejemplo, el buen humor ayuda. Las personas de buen humor son más cooperativas y propician la reciprocidad. Suelen obtener mejores resultados que los negociadores disgustados. Por supuesto, el éxito en las negociaciones también deja contentas a las personas, pero, en estos experimentos, su estado de ánimo no lo causa lo que ocurre en la negociación; es inducido antes de que negocien. Sin embargo, también los negociadores que pasan de un buen estado de ánimo a otro de enojo durante la negociación obtienen a menudo buenos resultados (algo que debemos recordar cuando nos enfrentamos a una persona obstinada).[15]

Por otro lado, un buen estado de ánimo nos hace más propensos a aceptar nuestras primeras impresiones como verdaderas sin cuestionarlas. En uno de los estudios de Forgas, los participantes leyeron un breve ensayo filosófico al que se adjuntó una imagen del autor.[16] Unos lectores vieron el estereotipo del profesor de Filosofía (hombre de mediana edad y con gafas). Otros vieron a una mujer joven. Es fácil adivinar que se trataba de una prueba de vulnerabilidad de los lectores a los estereotipos: ¿la gente

valora más el ensayo cuando se atribuye a un hombre de mediana edad que cuando creen que lo ha escrito una mujer joven? Lo hace, por supuesto; pero lo más importante es que la diferencia es mayor en estado de buen humor. Las personas que están de buen humor son más propensas a dejar que sus prejuicios afecten a su pensamiento.

Otros estudios comprobaron el efecto del estado de ánimo en la credulidad. Gordon Pennycook y sus colegas han realizado muchos estudios sobre las reacciones de la gente ante afirmaciones seudoprofundas sin sentido, creadas al juntar sustantivos y verbos seleccionados al azar de gurús populares en frases gramaticalmente correctas, como «La totalidad acalla infinitos fenómenos» o «El significado oculto transforma de modo incomparable la belleza abstracta».[17] La propensión a estar de acuerdo con tales afirmaciones es un rasgo conocido como *bullshit receptivity*. (La palabra *bullshit* [necedad, estupidez] se ha convertido en una especie de término técnico desde que Harry Frankfurt, filósofo de la Universidad de Princeton, publicó un esclarecedor libro, *On Bullshit*, en el que distinguía las necedades de otros tipos de imposturas.)[18]

Como era de esperar, algunas personas son más receptivas que otras a las necedades. Pueden dejarse impresionar por «afirmaciones aparentemente impresionantes que se presentan como verdaderas y significativas, pero que en realidad son vacuas».[19] Sin embargo, también aquí, esta credulidad no es una mera función de disposiciones permanentes e inmutables. Inducir el buen humor hace que las personas sean más receptivas a las necedades y más crédulas en general; son así menos aptas para detectar el engaño o identificar la información engañosa.[20] Por el contrario, testigos presenciales expuestos a una información engañosa son más capaces de ignorarla —y de evitar el falso testimonio— cuando están de mal humor.[21]

Incluso los juicios morales están fuertemente influidos por el estado de ánimo. En un estudio, los investigadores expusieron a los sujetos al dilema de la pasarela, un problema clásico de la filosofía moral.[22] En este experimento mental, cinco personas están a punto de morir atropelladas por un tranvía. Los sujetos deben imaginarse de pie en una pasarela debajo de la cual pronto pasará el tranvía. Deben decidir si empujan a un hombre grande desde la pasarela hasta las vías para que su cuerpo detenga el tranvía. Si lo hacen, se les dice, ese hombre morirá, pero las cinco personas se salvarán.

El dilema de la pasarela ilustra el conflicto entre los enfoques del razonamiento moral. El cálculo utilitario, asociado al filósofo inglés Jeremy Bentham, nos dirá que la pérdida de una vida es preferible a la pérdida de cinco. La ética deontológica, asociada a Immanuel Kant, prohíbe matar a alguien, incluso para salvar a otros. Este dilema contiene claramente un elemento destacado de emoción personal: empujar a un hombre desde un puente hacia la trayectoria de un tranvía que se aproxima es un acto especialmente repugnante. Tomar la decisión utilitaria de empujar al hombre requiere que las personas superen su aversión a un acto físicamente violento contra un extraño. Solo una minoría (en este estudio, menos de uno de cada diez) dice que lo haría.

Sin embargo, cuando los sujetos se encontraban en un estado de ánimo positivo —inducido por el visionado de una secuencia de vídeo de cinco minutos—, eran tres veces más propensos a decir que empujarían al hombre. Que consideremos el «no matarás» como un principio absoluto o que estemos dispuestos a matar a un desconocido para salvar a cinco personas debería reflejar nuestros valores más profundos. Sin embargo, nuestra elección parece depender del videoclip que acabamos de ver.

Hemos descrito estos estudios sobre el estado de ánimo con cierto detalle porque necesitamos subrayar una verdad importante: «Nadie es la misma persona en todo momento». Al variar el estado de ánimo (algo de lo que, por supuesto, somos conscientes), algunas características de la maquinaria cognitiva varían con él (algo de lo que «no» somos plenamente conscientes). Si se nos muestra un problema de juicio complejo, nuestro estado de ánimo en ese momento puede influir en nuestro planteamiento del problema y en las conclusiones a que lleguemos, incluso cuando creamos que nuestro estado de ánimo no tiene esa influencia y aunque podamos justificar con seguridad la respuesta que hemos dado. En suma, somos ruidosos.

Muchos otros factores secundarios inducen ruido de ocasión en los juicios. Entre los factores externos que no deberían influir en los juicios profesionales, pero que lo hacen, hay dos sospechosos principales: el estrés y el cansancio. Un estudio de casi setecientas mil consultas de atención primaria, por ejemplo, mostró que los médicos son muchísimo más propensos a recetar opioides al final de una larga jornada.[23] Sin duda, no hay ninguna razón por la que un paciente con una cita a las cuatro de la

tarde sufra más dolores que otro que se presenta a las nueve de la mañana. El hecho de que el médico vaya con retraso no debe influir en sus prescripciones. Y, de hecho, las prescripciones de otros tratamientos contra el dolor, como los antiinflamatorios no esteroideos y la derivación a fisioterapia, no muestran patrones similares. Cuando los médicos están presionados por el tiempo, parece que se inclinan más por una solución rápida, a pesar de sus serios inconvenientes. Otros estudios mostraron que, hacia el final de la jornada, los médicos son más propensos a recetar antibióticos[24] y menos a prescribir vacunas contra la gripe.[25]

Hasta el clima tiene una influencia medible en los juicios profesionales. Dado que estos juicios se realizan a menudo en salas con aire acondicionado, es probable que el efecto del clima esté «mediado» por el estado de ánimo (es decir, el clima no afecta directamente a las decisiones, sino que modifica el estado de ánimo de quien toma una decisión, que a su vez cambia su forma de decidir). El mal tiempo se asocia a la mejora de la memoria; las sentencias judiciales tienden a ser más severas cuando hace calor en el exterior; y los resultados de la bolsa se ven afectados por la luz solar.[26] En algunos casos, el efecto del tiempo es menos evidente. Uri Simonsohn demostró que los responsables de las admisiones universitarias prestan más atención a la cualificación académica de los candidatos en los días más nublados y son más sensibles a la cualificación no académica en los días más soleados. El título del artículo en el que informó de estos hallazgos es memorable: «Las nubes hacen que los empollones estén bien vistos».[27]

Otra fuente de variabilidad aleatoria en el juicio es el orden en que se examinan los casos. Cuando una persona considera un caso, las decisiones inmediatamente anteriores le sirven de marco de referencia implícito. Los profesionales que toman una serie de decisiones seguidas, incluidos jueces, prestamistas y árbitros de béisbol, se inclinan por restablecer una forma de equilibrio: tras una serie de decisiones que van en la misma dirección, es más probable que decidan en la dirección contraria a la que estaría estrictamente justificada. En consecuencia, los errores (y las injusticias) son inevitables. En Estados Unidos, los jueces de asilo político, por ejemplo, tienen un 19 por ciento menos de probabilidades de conceder asilo a un solicitante cuando los dos casos anteriores fueron admitidos. Una persona puede tener la suerte de que le concedan un préstamo si las dos solicitudes

anteriores fueron denegadas, incluso a esa misma persona se le habría denegado si las dos solicitudes anteriores hubieran sido aceptadas. Este comportamiento refleja un sesgo cognitivo conocido como la «falacia del jugador»: tendemos a subestimar la probabilidad de que sus buenas o malas rachas se produzcan por azar.[28]

EVALUACIÓN DEL RUIDO DE OCASIÓN

¿Cuál es la magnitud del ruido de ocasión relativa al ruido total del sistema? Aunque no hay un número único que se aplique a todas las situaciones, se da una regla general. Respecto a su magnitud, los efectos que hemos descrito en este capítulo son menores que las diferencias estables entre individuos en sus niveles y patrones de juicio.

Como hemos señalado, la posibilidad de que, por ejemplo, un solicitante de asilo sea admitido en Estados Unidos disminuye en un 19 por ciento si la audiencia se produce después de otras dos realizadas con éxito para el solicitante por el mismo juez. Esta variabilidad es ciertamente preocupante. Sin embargo, palidece en comparación con la variabilidad entre jueces: en un juzgado de Miami, Jaya Ramji-Nogales y sus coautores encontraron que un juez concedía el asilo al 88 por ciento de los solicitantes y otro solo al 5 por ciento.[29] (Se trata de datos reales, no de una auditoría del ruido, por lo que los solicitantes eran diferentes, pero se concedieron de forma casi aleatoria, y los autores comprobaron que las diferencias de país de origen no explicaban las discrepancias.) Ante tales disparidades, reducir una de estas cifras en un 19 por ciento no parece gran cosa.

Del mismo modo, los dactiloscopistas y los médicos a veces no están de acuerdo consigo mismos, pero con menos frecuencia que en desacuerdo con los demás. En todos los casos que revisamos en los que se pudo medir la parte de ruido de ocasión en el ruido total del sistema, el ruido de ocasión contribuyó menos que las diferencias entre individuos.

O, dicho de otro modo, uno no es siempre la misma persona, y es menos consistente en el tiempo de lo que cree. Aunque, para su tranquilidad, es más parecido a sí mismo ayer que a otra persona hoy.

RUIDO DE OCASIÓN: CAUSAS INTERNAS

Estado anímico, fatiga, clima, efectos de una secuencia... Muchos factores pueden desencadenar variaciones no deseadas en el juicio del mismo caso por parte de la misma persona. Podríamos esperar construir un escenario en el que se conozcan y controlen todos los factores externos que influyen en las decisiones. Al menos en teoría, este escenario debería reducir el ruido de ocasión, pero es probable que incluso este escenario no baste para eliminarlo por completo.

Michael Kahana y sus colegas de la Universidad de Pennsylvania estudian el rendimiento de la memoria. (La memoria no es una tarea de juicio según nuestra definición, pero es una tarea cognitiva de la que se pueden controlar rigurosamente las condiciones y medir fácilmente las variaciones de rendimiento.) En un estudio, pidieron a setenta y nueve sujetos que participaran en un análisis excepcionalmente exhaustivo del rendimiento de su memoria.[30] Los sujetos asistieron a veintitrés sesiones en días distintos, durante las cuales tuvieron que recordar palabras de veinticuatro listas diferentes de veinticuatro palabras cada una. El porcentaje de palabras recordadas define el rendimiento de la memoria.

Kahana y sus colegas no estaban interesados en las diferencias entre sujetos, sino en los indicadores de variabilidad en el rendimiento de cada sujeto. ¿Dependería el rendimiento de lo alerta que estuviese? ¿Del tiempo que durmiera la noche anterior? ¿De la hora del día? ¿Aumentaría su rendimiento con la práctica entre una sesión y la siguiente? ¿Se deterioraría en cada sesión a medida que se cansaran o aburrieran? ¿Serían unas listas de palabras más fáciles de memorizar que otras?

La respuesta a todas estas preguntas fue afirmativa, pero no por mucho. Un modelo que incorporara todos estos indicadores solo explicaba el 11 por ciento de la variación en el rendimiento de un determinado sujeto. Como dicen los investigadores, «nos sorprendió la variabilidad que seguía habiendo tras eliminar los efectos de nuestras variables de predicción». Incluso en este escenario tan controlado, los factores exactos que originaban el ruido de ocasión eran un misterio.

De todas las variables que estudiaron los investigadores, el indicador más potente del rendimiento de un sujeto en una lista particular no era un factor externo. El rendimiento en una lista de palabras lo predecía mejor lo bien que un sujeto lo había hecho en la lista inmediatamente anterior. A una lista recordada con éxito le seguía otra recordada con relativo éxito, y a una con

un recuerdo mediocre, otra con un rendimiento también mediocre. El rendimiento no varió aleatoriamente de una lista a otra: en cada sesión, hubo fluctuaciones con el paso del tiempo sin una causa externa evidente.

Estos hallazgos sugieren que el rendimiento de la memoria depende en gran parte, según Kahana y sus coautores, de «la eficiencia de los procesos neuronales endógenos que rigen la función de la memoria». En otras palabras, la variabilidad de la eficacia del cerebro de un momento a otro no se debe a influencias externas, como la meteorología o una distracción, sino que es una característica del propio funcionamiento de nuestro cerebro.

Es muy probable que la variabilidad intrínseca en el funcionamiento del cerebro también afecte a la calidad de nuestros juicios de un modo que no podemos esperar controlar. Esta variabilidad en el funcionamiento del cerebro debería hacer reflexionar a cualquiera que piense que se puede eliminar el ruido de ocasión. La comparación con el jugador de baloncesto en la línea de tiro libre no era tan simple como al principio pudo parecer: del mismo modo que los músculos del jugador nunca ejecutan exactamente el mismo movimiento, nuestras neuronas nunca funcionan exactamente igual. Si nuestra mente es un instrumento de medición, nunca será perfecto.

Sin embargo, podemos esforzarnos por controlar aquellas influencias indebidas que pueden ser controladas. Hacerlo es particularmente importante cuando los juicios se hacen en grupo, como veremos en el capítulo 8.

SOBRE EL RUIDO DE OCASIÓN

- «El juicio es como un lanzamiento libre: por mucho que intentemos repetirlo con precisión, nunca es exactamente idéntico.»
- «El juicio depende del estado de ánimo en que uno se encuentre, de los casos que acabe de discutir y hasta del tiempo que haga. No se es la misma persona en todo momento.»
- «Aunque uno no sea la misma persona que era la semana pasada, es menos diferente del que era la semana pasada que de otra persona hoy. El ruido de ocasión no es la mayor fuente de ruido del sistema.»

8
Cómo los grupos amplifican el ruido

El ruido en el juicio individual ya es bastante malo, pero la toma de decisiones en grupo añade otra capa al problema. Los grupos pueden ir en todo tipo de direcciones, dependiendo en parte de factores que tendrían que ser irrelevantes. Quién habla primero, quién habla el último, quién habla con seguridad en sí mismo, quién viste de negro, quién se sienta al lado de quién, quién sonríe, o frunce el ceño, o gesticula en el momento oportuno... Todos estos factores, y muchos más, afectan a los resultados. Todos los días, grupos similares toman decisiones muy diferentes, ya sea en cuestiones relacionadas con la contratación, la promoción, el cierre de oficinas, las estrategias de comunicación, las regulaciones ambientales, la seguridad nacional, las admisiones universitarias o el lanzamiento de nuevos productos.

Puede parecer extraño hacer hincapié en este punto, ya que en el capítulo anterior señalamos que la suma de los juicios de varios individuos reduce el ruido. Sin embargo, debido a la dinámica de los grupos, estos también pueden añadir ruido. Hay «multitudes sabias» cuyo juicio medio se acerca a la respuesta correcta, pero también hay multitudes que siguen a tiranos, que alimentan burbujas en el mercado, que creen en la magia o que están bajo el influjo de una ilusión compartida. Pequeñas diferencias pueden llevar a un grupo a decir «sí» con firmeza y a un grupo esencialmente idéntico a decir «no» con rotundidad. Y, debido a la dinámica entre los miembros del grupo —aquí nuestro énfasis—, el nivel de ruido puede ser elevado. Esta proposición es válida tanto si hablamos de ruido en grupos similares como de un único grupo cuyo juicio firme sobre un asunto importante debe considerarse como uno más en una nube de posibilidades.

EL RUIDO EN LA MÚSICA

Para probarlo, empezaremos en lo que podría parecer un lugar poco probable: el estudio a gran escala de Matthew J. Salganik y sus coautores

sobre las descargas de música.[1] Tal y como se preparó el estudio, los experimentadores crearon un grupo de control de miles de personas (visitantes de un sitio web moderadamente popular). Los miembros del grupo de control podían escuchar y descargar una o varias de las setenta y dos canciones de grupos nuevos. Las canciones tenían títulos muy gráficos: «Atrapado en una piel de naranja», «Roer», «Parche en el ojo», «Mago del béisbol v1» y «Agresión rosa». (Algunos de los títulos parecen directamente relacionados con lo que aquí nos ocupa: «Los mejores errores», «Yo soy el error», «La creencia por encima de la respuesta», «El misterio de la vida», «Deséame suerte» y «Fuera del bosque».)

A los participantes del grupo de control no se les dijo nada sobre lo que otros habían dicho o hecho. Se les dejaba que juzgaran por sí mismos qué canciones les gustaban y deseaban descargar. Sin embargo, Salganik y sus colegas también crearon otros ocho grupos, a los que se asignaron aleatoriamente otros miles de visitantes del sitio web. Para los miembros de esos grupos todo era igual, con una sola excepción: podían ver cuántas personas de su grupo particular habían descargado antes cada canción individual. Por ejemplo, si «Los mejores errores» era muy popular en un grupo, sus miembros lo verían, y también si nadie la descargaba.

Como los distintos grupos no diferían en ningún aspecto importante, el estudio se realizó básicamente en ocho ocasiones. Se podría predecir que, al final, las buenas canciones siempre subirían a la cima y las malas se hundirían hasta el fondo. De ser así, los distintos grupos terminarían con clasificaciones idénticas o, al menos, similares. Entre los grupos, no habría ruido. Y, de hecho, esa era en concreto la cuestión que Salganik y sus coautores pretendían elucidar. Se trataba de una prueba de un factor de ruido concreto: la «influencia social».

El hallazgo clave fue que las clasificaciones que hicieron los grupos eran muy dispares: entre los distintos grupos había gran cantidad de ruido. En un grupo, «Los mejores errores» podía ser un éxito espectacular, mientras que «Yo soy un error» podía fracasar. En otro grupo, «Yo soy un error» podía quedar muy bien, y «Los mejores errores» podía ser un desastre. Si una canción sacaba provecho de una aceptación inicial, podía quedar muy bien. Si no tenía esa ventaja, el resultado podía ser muy diferente.

Sin duda, las peores canciones (calificadas por el grupo de control) nunca acabaron en los primeros puestos, y las mejores nunca acabaron en los

últimos. Pero, por lo demás, podía pasar casi cualquier cosa. Como destacan los autores, «el nivel de éxito en la condición de influencia social era más imprevisible que en la condición independiente». En suma, las influencias sociales crean un ruido importante entre los grupos. Y, si se piensa en ello, se puede ver que los grupos individuales también fueron ruidosos, en el sentido de que su juicio a favor o en contra de una canción podía haber sido fácilmente distinto dependiendo de si se ganó una aceptación inicial.

Como más adelante demostraron Salganik y sus coautores, los resultados grupales pueden manipularse con bastante facilidad, porque la popularidad se refuerza a sí misma.[2] En un experimento de seguimiento un tanto diabólico, invirtieron las clasificaciones en el grupo de control (en otras palabras, mintieron sobre la popularidad de las canciones), con lo que la gente vio las canciones menos populares como las más populares, y viceversa. A continuación, los investigadores comprobaron lo que hacían los visitantes del sitio web. El resultado fue que la mayoría de las canciones antes impopulares se volvieron bastante populares, y la mayoría de las canciones populares quedaron muy mal. Dentro de los grupos más grandes, la popularidad y la impopularidad mostraron más de lo mismo después de que los investigadores engañaran a la gente sobre qué canciones eran populares. La única excepción fue que la canción más popular del grupo de control aumentó su popularidad con el tiempo, lo que significa que la clasificación invertida no pudo rebajar la de la considerada mejor canción. Sin embargo, en su mayor parte, la clasificación invertida ayudó a determinar la clasificación final.

Debería ser fácil apreciar cómo estos estudios retratan a grandes rasgos los juicios en grupo. Supongamos que un pequeño grupo formado por, digamos, diez personas está decidiendo si adopta alguna nueva y audaz iniciativa. Si uno o dos defensores hablan primero, pueden hacer que toda la audiencia se oriente en la dirección que prefieren. Lo mismo ocurre si los escépticos hablan primero. Al menos esto es así si las personas se influyen mutuamente, y en general lo hacen. Por esta razón, grupos por lo demás similares podrían acabar emitiendo juicios muy diferentes solo por el efecto del que habló primero e inició el equivalente a las primeras descargas. La popularidad de «Los mejores errores» y «Yo soy el error» tiene analogías en juicios profesionales de todo tipo. Y, si los grupos no escuchan algo semejante a los *rankings* de popularidad de esas canciones —un fuerte

entusiasmo, digamos, por esa audaz iniciativa—, la iniciativa podría no llegar a ninguna parte simplemente porque los que la apoyaban no dan voz a su opinión.

MÁS ALLÁ DE LAS DESCARGAS DE MÚSICA

Puede que un escéptico piense que el caso de las descargas de música es único, o al menos peculiar, y que nos dice poco sobre los juicios de otros grupos. Sin embargo, también se han hecho observaciones similares en muchas otras áreas.[3] Consideremos, por ejemplo, la popularidad de las propuestas de referéndum en el Reino Unido. A la hora de decidir si apoyan un referéndum, los ciudadanos deben juzgar, por supuesto, si es una buena idea, considerando todos los aspectos. Los patrones son parecidos a los observados por Salganik y sus coautores: un estallido inicial de popularidad se autorrefuerza y, si una propuesta obtiene pocos apoyos el primer día, está irremisiblemente condenada. En la política, como en la música, es mucho lo que depende de las influencias sociales y, en particular, de que la gente vea qué es lo que atrae o repele a otros.

Basándose directamente en el experimento de las descargas de música, el sociólogo Michael Macy, de la Universidad de Cornell, y sus colaboradores se preguntaron si las opiniones visibles de otras personas podían hacer que posiciones políticas identificables fueran populares entre los demócratas e impopulares entre los republicanos, o viceversa.[4] La respuesta breve fue un sí. Si demócratas de un grupo *online* veían que un punto de vista concreto estaba obteniendo una popularidad inicial entre los demócratas, respaldaban ese punto de vista, lo que finalmente hacía que la mayoría de los demócratas, en el grupo correspondiente, lo prefirieran. Pero si demócratas de otro grupo *online* veían que el mismo punto de vista estaba ganando una popularidad inicial entre los republicanos, rechazaban ese punto de vista, lo que finalmente hacía que la mayoría de los demócratas, en el grupo correspondiente, lo rechazaran. Los republicanos se comportaban de forma parecida. En suma, las posiciones políticas pueden ser como las canciones, en el sentido de que su destino final puede depender de la popularidad inicial. Como dicen los investigadores, «la variación fortuita en un pequeño número de personas que se adelantan» puede tener efectos importantes en la inclinación de grandes poblaciones (y conseguir que tanto

republicanos como demócratas acojan un conjunto de opiniones que, en realidad, no tienen nada que ver entre sí).

Consideremos ahora una cuestión que tiene que ver directamente con las decisiones de grupo en general: cómo juzga la gente los comentarios en los sitios web. Lev Muchnik, profesor de la Universidad Hebrea de Jerusalén, y sus colegas llevaron a cabo un experimento en una web que contaba distintas historias y permitía que la gente publicase comentarios que, a su vez, podían ser votados con el pulgar hacia arriba o hacia abajo.[5] Los investigadores otorgaron automática y artificialmente a ciertos comentarios de las historias un voto inmediato hacia arriba. Se podría pensar que, tras cientos de miles de visitas y valoraciones, un solo voto inicial sobre un comentario poco importaría. Una idea sensata, pero errónea. Después de ver un primer voto positivo (y recordemos que era totalmente artificial), el siguiente visitante era un 32 por ciento más propenso a dar un voto positivo.

Sorprendentemente, este efecto persistió en el tiempo. Al cabo de cinco meses, un solo voto inicial positivo aumentó artificialmente la valoración media de los comentarios en un 25 por ciento. El efecto de un único voto positivo anticipado es una receta para el ruido. Cualquiera que sea la razón de ese voto, puede producir un cambio a gran escala en la popularidad general.

Este estudio ofrece una pista sobre cómo cambian los grupos y por qué son ruidosos (de nuevo en el sentido de que grupos similares pueden hacer juicios muy diferentes, y grupos únicos pueden hacer juicios que solo son uno en una nube de posibilidades). Los miembros suelen estar en condiciones de ofrecer el equivalente funcional de un voto inicial a favor (o en contra) indicando acuerdo, neutralidad o desacuerdo. Si un miembro del grupo ha dado inmediatamente su aprobación, los demás miembros tienen motivos para hacerlo también. No cabe duda de que, cuando los grupos se inclinan por algunos productos, personas, movimientos e ideas, puede que no sea por sus méritos intrínsecos, sino por el equivalente funcional de los primeros votos favorables. Por supuesto, el propio estudio de Muchnik incluía grupos muy grandes. Pero lo mismo puede suceder en los pequeños, bien que de forma más espectacular, porque un voto inicial —a favor de algún plan, producto o veredicto— suele tener un gran efecto en los demás.

Hay un aspecto relacionado con esto último. Hemos hablado de la sabiduría de la multitud: si a un grupo numeroso de personas se le hace una

pregunta, hay muchas posibilidades de que la respuesta media se acerque mucho al objetivo. La suma de juicios puede ser una excelente manera de reducir tanto el sesgo como el error. Pero ¿qué sucede si la gente se escucha entre sí? Podemos pensar que el hecho de escuchar puede ayudar. Al fin y al cabo, las personas pueden aprender unas de otras y así averiguar qué es lo correcto. En circunstancias favorables, en las cuales la gente comparte lo que sabe, los grupos que deliberan pueden, en efecto, hacerlo bien. Sin embargo, la independencia es un requisito previo para la sabiduría de la multitud. Si la gente no se forma sus propios juicios y se basa en lo que piensan los demás, puede que después de todo las multitudes no sean tan sabias.

Las investigaciones han revelado exactamente este problema.[6] En tareas de estimación sencillas —el número de delitos en una ciudad, el aumento de la población en periodos específicos, la longitud de una frontera entre naciones—, las multitudes eran realmente sabias cuando registraban sus opiniones de forma independiente. Pero si conocían las estimaciones de otras personas —por ejemplo, la estimación media de un grupo de doce—, las multitudes lo hacían peor. Como dicen los autores, el problema son las influencias sociales, porque reducen «la diversidad del grupo sin disminuir el error colectivo». La ironía es que, aunque las múltiples opiniones independientes, debidamente sumadas, pueden ser sorprendentemente acertadas, solo un poco de influencia social puede conformar una especie de rebaño que socava la sabiduría de la multitud.

CASCADAS

Algunos de los estudios que estamos describiendo presuponen «cascadas de información». Estas son omnipresentes. Ayudan a explicar por qué grupos similares en las empresas, los gobiernos y otros ámbitos pueden ir en múltiples direcciones y por qué pequeños cambios pueden producir resultados tan diferentes y, por tanto, ruido. Solo podemos ver una historia tal y como se desarrolló realmente, pero, para muchos grupos y decisiones de grupo, hay nubes de posibilidades, de las cuales solo una se hace realidad.

Para saber cómo funcionan las cascadas informativas, imaginemos que diez personas están en una gran oficina decidiendo a quién contratar para un

puesto importante. Hay tres candidatos principales: Thomas, Sam y Julie. Supongamos que los miembros del grupo exponen sus opiniones de forma secuencial. Cada uno atiende suficientemente a las opiniones de los demás. Arthur es el primero en hablar. Sostiene que la mejor opción es Thomas. Barbara conoce ahora el parecer de Arthur; seguramente secundará su opinión si también está entusiasmada con Thomas. Pero supongamos que no está segura de quién es el mejor candidato. Si confía en Arthur, podría solo estar de acuerdo: Thomas es el mejor. Y, como confía bastante en Arthur, apoya su parecer.

Ahora pasemos a una tercera persona, Charles. Tanto Arthur como Barbara han dicho que quieren contratar a Thomas, pero la opinión de Charles, basada en lo que sabe que es poca información, es que Thomas no es la persona adecuada para el trabajo y que Julie es la mejor candidata. Aunque Charles tenga esa opinión, bien podría ignorar lo que sabe y seguir sin más a Arthur y a Barbara. Si lo hace, no será porque Charles se acobarde, sino más bien porque es un oyente respetuoso. Simplemente podría pensar que tanto Arthur como Barbara tienen pruebas que justifican su entusiasmo.

A menos que David piense que su propia información es realmente mejor que la de los que le precedieron, seguirá el ejemplo de estos. Si lo hace, David estará en una cascada. Es cierto que se resistirá si tiene motivos suficientes para pensar que Arthur, Barbara y Charles están equivocados. Sin embargo, si carece de esos motivos, es probable que les siga la corriente.

Es importante destacar que Charles o David pueden tener información o conocimientos sobre Thomas (o los otros candidatos) que Arthur y Barbara ignoran. Si esta «información privada» se hubiera compartido, podría haber cambiado las opiniones de Arthur o Barbara. Si Charles y David hubieran hablado primero, no solo habrían expresado su parecer sobre los candidatos, sino que habrían aportado información que podría haber influido en los demás participantes. Pero, al ser los últimos en hablar, su información privada podría seguir siendo privada.

Supongamos ahora que Erica, Frank y George deben expresar sus opiniones. Si Arthur, Barbara, Charles y David han dicho antes que Thomas es el mejor, cada uno de ellos podría decir lo mismo, aunque tengan buenas razones para pensar que otra opción sería mejor. Por supuesto, pueden

oponerse al creciente consenso si es obviamente erróneo. Pero ¿y si la decisión no está clara? El truco en este ejemplo es que el juicio inicial de Arthur ha puesto en marcha un proceso por el que varias personas se ven abocadas a participar en cascada, lo que lleva al grupo a optar de forma unánime por Thomas (aunque algunos de los que lo apoyan no tengan en realidad ninguna opinión y otros piensen que no es en absoluto la mejor opción).

Desde luego, este ejemplo es del todo artificial. Sin embargo, dentro de los grupos, del tipo que fueren, siempre sucede algo así. La gente aprende de otros, y, si a los primeros que hablan parece gustarles algo o quieren hacer determinadas cosas, los demás podrían asentir. Al menos esto es así si no tienen razones para desconfiar de ellos y si carecen de alguna buena razón para pensar que están equivocados.

Para nuestros fines, lo más importante es que las cascadas de información hacen posible, y hasta probable, el ruido en los grupos. En el ejemplo que hemos puesto, Arthur habló primero y prefirió a Thomas. Pero supongamos que Barbara hubiera hablado primero y se hubiese inclinado por Sam. O que Arthur hubiera intuido algo diferente y hubiese preferido a Julie. En suposiciones plausibles, el grupo habría preferido a Sam o a Julie, no porque fuesen mejores, sino porque así es como habría funcionado la cascada. Este es el hallazgo central del experimento de las descargas de música (y sus primos).

Obsérvese que no es necesariamente irracional que las personas participen en cascadas de información. Si no están seguras de a quién contratar, puede ser inteligente seguir a otros. A medida que aumenta el número de personas que comparten el mismo punto de vista, confiar en ellas se vuelve aún más inteligente. Sin embargo, hay dos problemas. En primer lugar, se tiende a ignorar la posibilidad de que la mayoría de las personas del colectivo se encuentren también en cascada y no estén haciendo juicios independientes por sí mismas. Cuando vemos que tres, diez o veinte personas se suman a alguna conclusión, es muy posible que subestimemos hasta qué punto todos siguen a sus predecesores. Podríamos pensar que su acuerdo compartido refleja la sabiduría colectiva, aunque refleje las opiniones iniciales de unas pocas personas. En segundo lugar, las cascadas de información pueden llevar a grupos de personas en direcciones

verdaderamente terribles. A fin de cuentas, Arthur podría haberse equivocado con Thomas.

La información no es, por supuesto, la única razón por la que los miembros del grupo influyen unos en otros. Las presiones sociales también son importantes. En una empresa, o en un Gobierno, sus integrantes pueden callarse para no parecer antipáticos, agresivos, obtusos o estúpidos. Quieren ser jugadores de equipo. Por eso siguen las opiniones y acciones de los demás. Creen que saben lo que es correcto, o probablemente correcto, pero siguen el visible consenso del grupo o las opiniones de los primeros en hablar para mantenerse bien integrados en el grupo.

Con pequeñas variaciones, la historia de la contratación que acabamos de contar puede seguir su curso, pero no porque las personas estén aprendiendo unas de otras cuáles puedan ser los méritos de Thomas, sino porque no quieren parecer antipáticas o necias. El juicio anticipado de Arthur a favor de Thomas podría iniciar una especie de efecto de arrastre, imponiendo finalmente una fuerte presión social sobre Erica, Frank o George, simplemente porque todos los demás han preferido a Thomas. Y lo mismo que sucede en las cascadas de información, ocurre también en las de presión social: las personas pueden exagerar la convicción de las que han hablado antes. Si apoyan a Thomas, puede que no lo hagan porque realmente prefieran a Thomas, sino porque alguien que habló antes, o alguien influyente, lo apoyó. Y, sin embargo, los miembros del grupo acaban sumando su voz al consenso y aumentando así el nivel de presión social. Este es un fenómeno conocido en empresas y en oficinas públicas, y puede conducir a la confianza y al respaldo unánime de un juicio que en sí sea erróneo.

En los grupos, las influencias sociales también producen ruido. Si alguien comienza una reunión inclinándose por un cambio importante en la dirección de la empresa, esa persona puede iniciar un debate que lleve al grupo a apoyar de manera unánime el cambio. Su acuerdo podría ser producto de las presiones sociales, no de la convicción. Si otra persona hubiera iniciado la reunión indicando un punto de vista diferente, o si la primera en hablar hubiera decidido permanecer en silencio, la discusión podría haber tomado una dirección completamente distinta (y por la misma razón). Grupos muy similares pueden acabar en posiciones discrepantes debido a las presiones sociales.

En Estados Unidos, y en muchos otros países, los casos penales (y muchos civiles) suelen ser juzgados por un jurado. Uno esperaría que, en sus deliberaciones, los jurados tomasen decisiones más sabias que los individuos que componen estos órganos de deliberación. Sin embargo, el estudio de los jurados descubre un tipo distinto de influencia social que también es una fuente de ruido: la «polarización de grupo». La idea básica es que, cuando las personas hablan entre sí, suelen terminar en un punto más extremo en la línea de sus primeras inclinaciones. Si, por ejemplo, la mayoría de las personas de un grupo de siete tienden a pensar que abrir una nueva oficina en París sería una buena idea, es probable que el grupo concluya, tras el debate, que abrir esa oficina sería una fantástica idea. Las discusiones internas suelen crear mayor confianza, mayor unidad y mayor extremismo, a menudo en la forma de un redoblado entusiasmo. La polarización de los grupos no solo se produce en los jurados; cuando los grupos emiten juicios profesionales, también suelen polarizarse.

En una serie de experimentos, estudiamos las decisiones de los jurados cuando imponen indemnizaciones por daños y perjuicios en casos de responsabilidad de fabricantes de ciertos productos. La decisión de cada jurado es penalizar a la empresa con una cantidad monetaria por su conducta punible y ser un elemento disuasorio para otras. (Volveremos sobre estos estudios y los describiremos con más detalle en el capítulo 15.) Para nuestro propósito, consideremos un experimento que compara jurados deliberantes del mundo real y «jurados estadísticos».[7] Empezamos presentando historias de casos a los ochocientos noventa y nueve participantes de nuestro estudio, y les pedimos que hicieran sus propios juicios independientes sobre ellos utilizando una escala de siete grados para expresar su indignación e intención punitiva y una escala en dólares para las indemnizaciones (si las hubiera). A continuación, utilizamos estas respuestas individuales para crear, con la ayuda de una computadora, millones de jurados estadísticos, es decir, grupos virtuales de seis personas (reunidas al azar). En cada jurado estadístico, tomamos la mediana de los seis juicios individuales como veredicto.

Encontramos, en suma, que los juicios de estos jurados estadísticos eran mucho más coherentes. El ruido se redujo sustancialmente. El ruido

reducido era un efecto mecánico de la suma estadística: el ruido presente en los juicios independientes, individuales, siempre se reduce al promediarlos.

Sin embargo, los jurados del mundo real no son jurados estadísticos; se reúnen y discuten sus puntos de vista sobre el caso. Es razonable preguntarse si los jurados que deliberan tienden, de hecho, al juicio de sus miembros que representan a la mediana. Para averiguarlo, enlazamos el primer experimento con otro nuevo, este con más de tres mil ciudadanos aptos para ser miembros de jurados y más de quinientos jurados de seis personas.[8]

Los resultados fueron claros. En el mismo caso, los jurados deliberantes eran mucho más ruidosos que los estadísticos, un claro reflejo del ruido de la influencia social. La deliberación tuvo el efecto de aumentar el ruido.

Hubo otro hallazgo interesante. Cuando el miembro que representaba a la mediana de un grupo de seis personas estaba solo moderadamente indignado y era partidario de una penalización indulgente, el veredicto del jurado que deliberaba solía ser aún más indulgente. Cuando, por el contrario, el miembro que representaba a la mediana de un grupo de seis personas estaba indignado y expresaba una seria intención punitiva, el jurado que deliberaba solía acabar más indignado y siendo aún más severo. Y, cuando esta indignación se expresaba en forma de indemnización, se observaba una tendencia sistemática a las indemnizaciones superiores a las del miembro del jurado que representaba a la mediana. De hecho, el 27 por ciento de los jurados eligió una indemnización tan alta (o incluso más) como la de su miembro más severo. Los jurados deliberantes no solo eran más ruidosos que los estadísticos, sino que también acentuaban las opiniones de los individuos que los componían.

Recordemos el descubrimiento básico de la polarización de grupos: las personas, después de hablar entre ellas, suelen acabar en un punto más extremo en la línea de sus primeras inclinaciones. Nuestro experimento ilustra este efecto. Los jurados deliberantes experimentaron un cambio hacia una mayor indulgencia (cuando el miembro que representaba a la mediana era indulgente) y un cambio hacia una mayor severidad (cuando ese miembro era severo). Del mismo modo, los jurados inclinados a imponer indemnizaciones acabaron imponiendo penas más severas que las que los miembros de la mediana habrían preferido.

Las explicaciones de la polarización de grupos son, a su vez, similares a las de los efectos de cascada. La información desempeña un papel fundamental. Si una mayoría está a favor de un castigo severo, el grupo escuchará muchos argumentos a favor del castigo severo y menos argumentos en contra. Si los miembros del grupo se escuchan unos a otros, se desplazarán en la dirección de la tendencia dominante, lo que hará que el grupo esté más unido, se sienta más seguro y sea más extremo. Y, si las personas se preocupan por su reputación dentro del grupo, de igual modo se desplazarán en la dirección de la tendencia dominante, lo que también generará una polarización.

La polarización de los grupos puede, por supuesto, producir errores. Y a menudo lo hace, pero nuestro foco de atención es aquí la variabilidad. Como hemos visto, una suma de juicios reducirá el ruido, y, para este fin, cuantos más juicios, mejor. Por eso los jurados estadísticos son menos ruidosos que los miembros individuales. Al mismo tiempo, comprobamos que los jurados deliberantes son más ruidosos que los estadísticos. Cuando grupos con una situación similar acaban difiriendo, la razón suele ser la polarización de grupo. Y el ruido resultante puede ser muy fuerte.

En las empresas, en los gobiernos y en cualquier otro ámbito, las cascadas y la polarización pueden dar lugar a grandes disparidades entre grupos que analizan el mismo problema. La potencial dependencia del juicio de unos pocos individuos —los que hablan primero o los que tienen mayor influencia— debería ser especialmente preocupante ahora que hemos puesto de manifiesto lo ruidosos que pueden ser los juicios individuales. Hemos visto que el ruido de nivel y el ruido de patrón hacen que las diferencias entre las opiniones de los miembros del grupo sean mayores de lo que deberían ser (y mayores de lo que esperaríamos). También hemos visto que el ruido de ocasión —cansancio, estado de ánimo, comparaciones — puede afectar al juicio de la primera persona que habla. Y que la dinámica del grupo puede amplificar ese ruido. La consecuencia es que los grupos que deliberan tienden a ser más ruidosos que los grupos estadísticos, que se limitan a promediar los juicios individuales.

Dado que muchas de las decisiones más importantes en las empresas y en los gobiernos se toman después de algún proceso de deliberación, es particularmente importante estar alerta ante este riesgo. Las organizaciones y sus dirigentes deben tomar medidas para controlar el ruido en los juicios

individuales de sus miembros. También han de gestionar los grupos de deliberación de forma que se reduzca el ruido, no que se amplifique. Las estrategias de reducción del ruido que propondremos tratan de conseguir este objetivo.

SOBRE LAS DECISIONES DE GRUPO

- «Todo parece depender de la aceptación inicial. Será mejor que trabajemos duro para que nuestro nuevo lanzamiento tenga una primera semana estupenda.»
- «Como siempre sospeché, las ideas sobre política y economía se parecen mucho a las estrellas de cine. Si la gente piensa que a los demás les gustan, esas ideas pueden llegar lejos.»
- «Siempre me ha preocupado que, cuando mi equipo se reúna, acabemos confiados y unidos, y firmemente comprometidos con el curso de la acción que elijamos. Me temo que hay algo en nuestros procesos internos que no marcha del todo bien.»

TERCERA PARTE

El ruido en los juicios predictivos

Muchos juicios son predicciones y, como las predicciones verificables pueden evaluarse, podemos aprender mucho sobre el ruido y el sesgo estudiándolas. En esta parte del libro, nos centraremos en los juicios predictivos.

El capítulo 9 compara la exactitud de las predicciones realizadas por profesionales, por máquinas y por medio de reglas simples. A nadie le sorprenderá nuestra conclusión de que los profesionales son los terceros en esta competición. En el capítulo 10 analizamos las razones de este resultado y demostramos que el ruido es un factor importante en la inferioridad del juicio humano.

Para llegar a estas conclusiones, debemos evaluar la calidad de las predicciones, y para ello necesitamos una medida de la exactitud de estas, una forma de responder a esta pregunta: ¿en qué medida «covarían» las predicciones con los resultados? Si un departamento de recursos humanos califica rutinariamente el potencial de los nuevos contratos, por ejemplo, podemos esperar unos años para saber cómo actúan los empleados y ver en qué medida las calificaciones del potencial covarían con las evaluaciones de su rendimiento. Las predicciones son exactas en la medida en que los empleados cuyo potencial fue calificado como alto cuando fueron contratados también obtienen valoraciones altas de su trabajo.

Una medida que recoge esta intuición es el «porcentaje de concordancia» (PC), que responde a una pregunta más específica. Supongamos que elegimos a un par de empleados al azar. ¿Cuál es la probabilidad de que el que ha obtenido una puntuación más alta en una evaluación de potencial muestre también un mejor rendimiento en el trabajo?[1] Si la exactitud de las primeras calificaciones fuese perfecta, el PC sería del cien por cien: la clasificación de los empleados por potencial sería una predicción perfecta de su clasificación final por rendimiento. Si las predicciones fuesen totalmente inútiles, la concordancia se produciría solo por azar, y el empleado con «mayor potencial» tendría la misma probabilidad de rendir más que de rendir menos: el PC sería del 50 por ciento. En el capítulo 9 analizaremos este ejemplo, que ha sido estudiado en profundidad. Para

poner uno más sencillo, el PC para la longitud de los pies y la estatura en hombres adultos es del 71 por ciento. Si miramos a dos hombres, primero a la cabeza y luego a los pies, hay un 71 por ciento de posibilidades de que el más alto de los dos tenga también los pies más grandes.

El PC es una medida de covariación intuitiva inmediata, lo que supone una gran ventaja, pero no es la medida estándar que utilizan los científicos sociales. La medida estándar es el «coeficiente de correlación» (r), que varía entre 0 y 1 cuando dos variables están relacionadas positivamente. En el ejemplo anterior, la correlación entre la estatura y el tamaño de los pies es de, aproximadamente, .60.[2]

Hay muchas maneras de pensar en el coeficiente de correlación. He aquí una bastante intuitiva: la correlación entre dos variables es su porcentaje de determinantes compartidos. Imaginemos, por ejemplo, que un rasgo está totalmente determinado por la genética. Esperaríamos encontrar una correlación de .50 en ese rasgo entre hermanos que tienen el 50 por ciento de sus genes en común, y una correlación de .25 entre primos hermanos, que tienen el 25 por ciento de sus genes en común. También podemos interpretar una correlación de .60 entre la estatura y el tamaño de los pies como una indicación de que el 60 por ciento de los factores causales que determinan la estatura condicionan también el tamaño de los zapatos.

Las dos medidas de covariación que hemos descrito están directamente relacionadas entre sí. La tabla 1 presenta el PC para varios valores del coeficiente de correlación.[3] En el resto de este libro, siempre presentamos las dos medidas juntas cuando hablamos del rendimiento de humanos y de modelos.

En el capítulo 11 discutimos un importante límite en la exactitud de la predicción: el hecho de que la mayoría de los juicios se hagan en un estado de lo que llamamos «ignorancia objetiva», porque muchas cosas de las que depende el futuro no se pueden conocer.

Coeficiente de correlación	Porcentaje de concordancia (PC)
.00	50%
.10	53%
.20	56%
.30	60%
.40	63 %
.60	71%
.80	79%
1,00	100%

TABLA 1. Coeficiente de correlación y porcentaje de concordancia (PC)

Sorprendentemente, la mayor parte del tiempo nos las arreglamos para permanecer ajenos a esta limitación y hacer predicciones con confianza (o, más bien, exceso de confianza). Por último, en el capítulo 12, mostramos que la ignorancia objetiva no solo afecta a nuestra capacidad de predecir acontecimientos, sino incluso a nuestra capacidad de comprenderlos (una parte importante de la respuesta al enigma de por qué el ruido tiende a ser invisible).

9
Juicios y modelos

Es muy común el interés en predecir el futuro rendimiento de las personas en el trabajo, el de uno mismo y el de otras. La predicción del rendimiento es, por tanto, un ejemplo útil de juicio profesional predictivo. Pensemos, por ejemplo, en dos ejecutivas de una gran empresa. Monica y Nathalie fueron evaluadas por una firma consultora especializada en el momento de su contratación y recibieron calificaciones en una escala de 1 a 10 en cuanto a liderazgo, comunicación, habilidades interpersonales, habilidades técnicas relacionadas con el trabajo y motivación para el siguiente puesto (tabla 2). Nuestra tarea consiste en predecir las evaluaciones de su rendimiento dos años después de su contratación utilizando también una escala de 1 a 10.

	Liderazgo	Comunicación	Habilidades interpersonales	Habilidades técnicas	Motivación	Nuestra predicción
Monica	4	6	4	8	8	
Nathalie	8	10	6	7	6	

TABLA 2. Dos candidatas a un puesto ejecutivo

La mayoría de la gente, cuando se enfrenta a este tipo de problema, se limita a echar un vistazo a cada línea y emitir un juicio rápido, a veces después de calcular mentalmente la media de las puntuaciones. Si solo hace eso, probablemente concluirá que Nathalie es la candidata más fuerte y que la diferencia entre ella y Monica es de 1 o 2 puntos.

¿JUICIO O FÓRMULA?

El enfoque informal adoptado para este problema se conoce como «juicio clínico». Considerada la información, se hace un cálculo rápido, se consulta

a la intuición y se emite un juicio. De hecho, el juicio clínico es el proceso que hemos descrito como juicio en este libro.

Supongamos ahora que alguien realiza la tarea de predicción como participante en un experimento. Monica y Nathalie proceden de una base de datos de varios cientos de directivos contratados hace algunos años que recibieron calificaciones en cinco dimensiones distintas. Ha utilizado estas calificaciones para predecir el éxito de los directivos en el trabajo. Ahora están disponibles las evaluaciones de su rendimiento en sus nuevas funciones. ¿En qué medida aquellas se ajustan a sus juicios clínicos sobre su potencial?

Este ejemplo está basado libremente en un estudio real de predicción del rendimiento.[1] Quien haya participado en ese estudio, probablemente no estará satisfecho con sus resultados. Los psicólogos con un nivel de doctorado, contratados por una empresa consultora internacional para hacer tales predicciones, obtuvieron una correlación de .15 con las evaluaciones de rendimiento (PC = 55%). En otras palabras, cuando calificaron a un candidato como más fuerte que otro —como lo hicieron con Nathalie y Monica—, la probabilidad de que su candidata preferida terminara con una calificación de rendimiento más alta era del 55 por ciento, apenas mejor que el azar. No es, por tanto, un resultado impresionante.

Tal vez se piense que la exactitud fue escasa porque las calificaciones que se mostraron eran inútiles para una predicción. Entonces debemos preguntarnos: ¿cuánta información predictiva útil contienen realmente las calificaciones de las candidatas? ¿Cómo se pueden combinar en una puntuación predictiva que tenga la mayor correlación posible con el rendimiento?

Un método estadístico estándar responderá a estas preguntas. En el presente estudio, se obtiene una correlación óptima de .32 (PC = 60 %), lejos todavía de ser impresionante, pero mucho más alta que la que obtuvieron las predicciones clínicas.

Esta técnica, denominada «regresión múltiple», produce una puntuación predictiva que es una media ponderada de los predictores.[2] Encuentra el conjunto óptimo de ponderaciones, elegido para maximizar la correlación entre la predicción compuesta y la variable objetivo. Las ponderaciones óptimas minimizan el ECM (error cuadrático medio) de las predicciones, lo que constituye un buen ejemplo del papel dominante del principio de los

mínimos cuadrados en estadística. Como se puede esperar, el predictor que está más correlacionado con la variable objetivo obtiene una gran ponderación, y los predictores inútiles obtienen una ponderación cero.[3] Las ponderaciones también podrían ser negativas: el número de multas de tráfico no pagadas de la candidata probablemente obtendría una ponderación negativa como predictor de su éxito como directiva.

El uso de la regresión múltiple es un ejemplo de «predicción mecánica». Hay muchos tipos de predicción mecánica, desde reglas sencillas («contrata a cualquiera que haya terminado los estudios secundarios») hasta sofisticados modelos de inteligencia artificial. Sin embargo, los modelos de regresión lineal son los más comunes (se les ha llamado «el caballo de batalla de la investigación sobre el juicio y la toma de decisiones».)[4] Para minimizar la jerga, nos referiremos a los modelos lineales como «modelos simples».

El estudio que ilustramos con Monica y Nathalie fue una de las muchas comparaciones de predicciones clínicas y mecánicas, que comparten una estructura sencilla:[5]

- Se utiliza un conjunto de «variables predictoras» (en nuestro ejemplo, las calificaciones de los candidatos) para predecir un «resultado final» (las evaluaciones del trabajo de las mismas personas).
- Los juzgadores humanos hacen «predicciones clínicas».
- Una regla (como la regresión múltiple) utiliza los mismos predictores para generar «predicciones mecánicas» de los mismos resultados.
- Se compara la exactitud global de las predicciones clínicas y mecánicas.

MEEHL: EL MODELO ÓPTIMO NOS GANA

Cuando se enseña a la gente la predicción clínica y la mecánica, esta quiere saber cómo se comparan ambas. ¿Cómo es de bueno el juicio humano frente a una fórmula?

La cuestión ya se había planteado antes, pero no atrajo mucha atención hasta 1954, cuando Paul E. Meehl, profesor de Psicología de la Universidad de Minnesota, publicó un libro titulado *Clinical Versus Statistical Prediction. A Theoretical Analysis and a Review of the Evidence*.[6] Meehl revisó veinte estudios en los que se comparaba un juicio clínico con una predicción mecánica para resultados como el éxito académico y el pronóstico psiquiátrico. Llegó a la firme conclusión de que las reglas

mecánicas simples eran generalmente superiores al juicio humano. Meehl descubrió que los doctores y otros profesionales son lamentablemente débiles en lo que a menudo consideran su único punto fuerte: la capacidad de asimilar información.

Para apreciar lo sorprendente de este hallazgo, así como su relación con el ruido, hay que entender cómo funciona un simple modelo de predicción mecánica. Su característica determinante es que se aplica la misma regla a todos los casos. Cada predictor tiene un peso, y ese peso no varía de un caso a otro. Podría pensarse que esta estricta limitación pone a los modelos en gran desventaja con respecto a los juzgadores humanos. En nuestro ejemplo, puede que alguien haya pensado que la combinación de motivación y habilidades técnicas en Monica sería un valor importante y compensaría sus limitaciones en otras áreas. Y quizá puede que también alguien haya pensado que las debilidades de Nathalie en estas dos áreas no serían un serio problema, dados sus otros puntos fuertes. Implícitamente imaginaría distintas vías hacia el éxito para las dos mujeres. Estas especulaciones clínicas plausibles asignan en efecto diferentes pesos a los mismos predictores en los dos casos (una sutileza que está fuera del alcance de un modelo simple).

Otra restricción del modelo simple es que un aumento de 1 unidad en un predictor siempre produce el mismo efecto (y la mitad del efecto de un aumento de 2 unidades). La intuición clínica a menudo infringe esta norma. Si, por ejemplo, a alguien le impresionó el 10 perfecto de Nathalie en habilidades de comunicación y decidió que esa puntuación suponía un impulso a su predicción, hizo algo que un simple modelo no hará. En una fórmula de media ponderada, la diferencia entre una puntuación de 10 y otra de 9 debe ser la misma que la diferencia entre un 7 y un 6. El juicio clínico no obedece a esa norma. En cambio, refleja la intuición común de que la misma diferencia puede no tener consecuencias en un contexto y ser crítica en otro. Tal vez ese alguien quiera comprobarlo, pero sospechamos que ningún modelo simple podría explicar con exactitud sus juicios sobre Monica y Nathalie.

El estudio que utilizamos para estos casos era un claro ejemplo del patrón de Meehl. Como hemos señalado, las predicciones clínicas obtuvieron una correlación de .15 (PC = 55 %) con el rendimiento laboral, pero la predicción mecánica obtuvo una correlación de .32 (PC = 60 %). Pensemos

en la confianza que al juzgador clínico le inspiraron los méritos relativos de los casos de Monica y Nathalie. Los resultados de Meehl indican con claridad que cualquier satisfacción que el juzgador haya sentido con la calidad de su juicio era una ilusión: la «ilusión de validez».

La ilusión de validez se da en todos los casos en que se hacen juicios predictivos debido a un fallo común a la hora de distinguir entre dos etapas de la tarea de predicción: la evaluación de los casos según las pruebas disponibles y la predicción de los resultados reales. A menudo se puede estar bastante seguro de cuál de las dos candidatas «parece» mejor, pero estimar cuál de ellas «es» realmente mejor es harina de otro costal. Es seguro afirmar, por ejemplo, que Nathalie parece una candidata mejor que Monica, pero no es en absoluto seguro afirmar que Nathalie será una ejecutiva con más éxito que Monica. La razón es sencilla: se conoce la mayor parte de lo que hay que saber para evaluar los dos casos, pero la mirada al futuro es muy incierta.

Por desgracia, la diferencia se difumina en nuestro pensamiento. Quien se siente confundido con la distinción entre casos y predicciones está en inmejorable compañía: todo el mundo encuentra esa distinción confusa. Sin embargo, si está tan seguro de sus predicciones como de su evaluación de los casos, es víctima de la ilusión de validez.

Los doctores no son inmunes a la ilusión de validez. Podemos imaginar la respuesta de los psicólogos clínicos al hallazgo de Meehl de que las fórmulas triviales, aplicadas con coherencia, superan al juicio clínico. Su reacción combinó la conmoción, la incredulidad y el desprecio por una investigación superficial que pretendía estudiar las maravillas de la intuición clínica. Esta reacción es fácil de entender: la idea de Meehl contradice la experiencia subjetiva del juicio, y la mayoría de nosotros confiamos en nuestra experiencia antes que en las afirmaciones de un estudioso.

Sin embargo, Meehl se mostró ambivalente en relación con sus hallazgos. Como su nombre ha quedado asociado a la superioridad de la estadística sobre el juicio clínico, lo imaginaríamos como un crítico implacable de la perspicacia humana, o como el padrino de los *quants* (analistas cuantitativos), como diríamos hoy; eso, no obstante, sería una caricatura. Aparte de su carrera académica, Meehl era un psicoanalista en activo. Una foto de Freud colgaba en su despacho.[7] Era un erudito que no solo

impartía clases de Psicología, sino también de Filosofía y Derecho, y escribió sobre metafísica, religión, politología y hasta parapsicología.[8] (Insistía en que «algo hay en la telepatía».) Ninguna de estas características encaja con el estereotipo de un tipo duro pegado a los números. Meehl no tenía nada contra los médicos, ni mucho menos. Sin embargo, como él mismo dijo, las pruebas de la ventaja del enfoque mecánico para combinar *inputs* eran «muy numerosas y consistentes».[9]

«Muy numerosas y consistentes» es una descripción justa. Una revisión del año 2000 de ciento treinta y seis estudios confirmó de forma inequívoca que la suma mecánica supera al juicio clínico.[10] La investigación analizada en el artículo abarcaba una amplia variedad de temas, como el diagnóstico de ictericia, la aptitud para el servicio militar y la satisfacción conyugal. La predicción mecánica fue más precisa en sesenta y tres de los estudios, se declaró un empate estadístico en otros sesenta y cinco y la predicción clínica ganó la contienda en ocho casos. Estos resultados restan importancia a las ventajas de la predicción mecánica, que también es más rápida y barata que el juicio clínico. Además, los juzgadores humanos tenían una ventaja injusta en muchos de estos estudios, porque tenían acceso a información «privada» que no se suministraba al modelo informático.[11] Los resultados apoyan una conclusión contundente: los «modelos simples ganan a los humanos».

GOLDBERG: EL MODELO DE UNO MISMO LE GANA A UNO MISMO

La conclusión de Meehl suscita importantes cuestiones. ¿Por qué, exactamente, la fórmula es superior? ¿Qué hace mejor la fórmula? Una pregunta mejor sería la de qué es lo que los humanos hacen peor. La respuesta es que las personas son inferiores a los modelos estadísticos en muchos aspectos. Una de sus graves deficiencias es que son ruidosas.

Para respaldar esta conclusión, recurriremos a otra corriente de investigación sobre modelos simples que comenzó en la pequeña ciudad de Eugene (Oregón). Paul Hoffman era un psicólogo rico y visionario que se impacientaba con el mundo académico. Fundó un instituto de investigación donde reunió bajo un mismo techo a unos cuantos investigadores extraordinariamente eficaces, que convirtieron Eugene en un centro de fama mundial para el estudio del juicio humano.

Uno de estos investigadores fue Lewis Goldberg, más conocido por su papel protagonista en el desarrollo del modelo de los cinco grandes de la personalidad. A finales de la década de 1960, siguiendo los trabajos anteriores de Hoffman, Goldberg estudió modelos estadísticos que describen los juicios de un individuo.[12]

Es tan fácil construir un modelo de un juzgador como construir un modelo de la realidad. Se utilizan los mismos predictores. En nuestro ejemplo inicial, los predictores son las cinco calificaciones de la capacidad de un directivo. Y se utiliza la misma herramienta: la regresión múltiple. La única diferencia es la variable objetivo. En lugar de predecir un conjunto de resultados reales, la fórmula se aplica para predecir un conjunto de juicios (por ejemplo, los juicios que «uno» haría sobre Monica, Nathalie y otros directivos).

La idea de modelar los propios juicios individuales como una media ponderada puede parecer totalmente extraña, porque no es así como uno se forma sus opiniones. Cuando se piensa clínicamente en Monica y Nathalie, no se aplica la misma regla a ambos casos. De hecho, no se aplica ninguna regla. El modelo del juzgador no es una descripción realista de cómo en verdad juzga un juzgador.

Sin embargo, aunque uno no calcule realmente una fórmula lineal, puede hacer sus juicios «como si» la calculase. Los jugadores de billar expertos actúan como si hubieran resuelto las complejas ecuaciones que describen la mecánica de un tiro concreto, aunque no estén haciendo nada de eso.[13] Del mismo modo, uno podría estar generando predicciones como si utilizara una fórmula sencilla (aunque lo que en realidad hace es mucho más complejo). Un modelo «como si» que predice lo que las personas harán con una exactitud razonable es útil, incluso cuando es obviamente erróneo como descripción del proceso. Este es el caso de los modelos simples de juicio. Una revisión exhaustiva de los estudios sobre el juicio encontró que, en doscientos treinta y siete estudios, la correlación media entre el modelo del juzgador y los juicios clínicos del juzgador era de .80 (PC = 79 %). Aunque dista mucho de ser perfecta, esta correlación es lo bastante alta como para respaldar una teoría del «como si».[14]

La cuestión que impulsó la investigación de Goldberg fue la de saber hasta qué punto un modelo sencillo del juzgador podría predecir los resultados reales. Dado que el modelo es una aproximación tosca al

juzgador, podríamos suponer de forma sensata que no puede funcionar tan bien. ¿Cuánta exactitud se pierde cuando el modelo sustituye al juzgador?

La respuesta puede resultar sorprendente. Las predicciones no perdieron exactitud cuando el modelo generó predicciones, sino que mejoraron. En la mayoría de los casos, el modelo superó al profesional en el que se basaba. El sucedáneo era mejor que el producto original.

Esta conclusión ha sido confirmada por estudios realizados en muchos campos. Una de las primeras réplicas del trabajo de Goldberg se refería a la previsión del éxito en los estudios de posgrado.[15] Los investigadores pidieron a noventa y ocho participantes que predijeran las calificaciones medias de noventa estudiantes a partir de diez pistas. Sobre la base de estas predicciones, los investigadores construyeron un modelo lineal de los juicios de cada participante y compararon la precisión con la que los participantes y los modelos de los participantes predecían la calificación media. En cada uno de los noventa y ocho participantes, el modelo lo hizo mejor que el participante. Décadas después, una revisión de cincuenta años de investigación concluyó que los modelos de los juzgadores superaban sistemáticamente a los juzgadores que aquellos modelaban.[16]

No sabemos si los participantes en estos estudios recibieron información personal sobre su rendimiento. Sin embargo, podemos imaginar la consternación de un juzgador si alguien le dijera que un tosco modelo de sus juicios —casi una caricatura— es en realidad más exacto que él. Para la mayoría de nosotros, la actividad del juicio es compleja, rica e interesante justamente porque no se ajusta a reglas simples. Nos sentimos mejor con nosotros mismos y con nuestra capacidad para emitir juicios cuando inventamos y aplicamos reglas complejas o tenemos una idea que hace que un caso individual sea diferente de otros; en suma, cuando emitimos juicios que no son reducibles a una simple operación de media ponderada. Los estudios sobre el modelo del juzgador refuerzan la conclusión de Meehl de que la sutileza es en gran medida un desperdicio. La complejidad y la riqueza no suelen conducir a predicciones más exactas.

¿Por qué esto es así? Para entender el hallazgo de Goldberg, tenemos que comprender a qué se deben las diferencias entre uno y el modelo de uno. ¿Cuál es la causa de las discrepancias entre nuestros juicios reales y los resultados de un modelo simple que los predice?

Un modelo estadístico de nuestros juicios no puede añadir nada a la información que contienen. Todo lo que el modelo puede hacer es restar y simplificar. En particular, el modelo simple de nuestros juicios no representará ninguna regla compleja que sigamos de forma sistemática. Si alguien piensa que la diferencia entre un 10 y un 9 en una calificación de habilidades comunicativas es más significativa que la diferencia entre un 7 y un 6, o que un candidato equilibrado que obtiene un sólido 7 en todas las dimensiones es preferible a otro que alcanza la misma media con claros puntos fuertes y marcados puntos débiles, el modelo de él no reproducirá sus complejas reglas (aunque las aplique con impecable coherencia).

Si nuestras sutiles reglas no se reproducen, perderemos exactitud cuando nuestra sutileza sea válida. Supongamos, por ejemplo, que hay que predecir el éxito en una tarea difícil a partir de dos datos, la habilidad y la motivación. Una media ponderada no es una buena fórmula, porque ninguna motivación es suficiente para superar un grave déficit en habilidades, y viceversa. Si utilizamos una combinación más compleja de los *inputs*, nuestra exactitud predictiva mejorará y será mayor que la conseguida por un modelo que no capte esa sutileza. Por otro lado, las reglas complejas a menudo crean la ilusión de validez y, de hecho, perjudican la calidad de nuestros juicios. Algunas sutilezas son válidas, pero muchas no lo son.

Además, un modelo simple de uno mismo no representará el patrón de ruido de sus juicios. No puede reproducir los errores positivos y negativos que resultan de las reacciones arbitrarias que uno pueda tener a un caso concreto. El modelo tampoco captará las influencias del contexto fugaz y de su estado mental cuando emita un juicio concreto. Lo más probable es que estos ruidosos errores de juicio no estén sistemáticamente correlacionados con nada, lo que significa que, para la mayor parte de los fines, pueden considerarse aleatorios.

El efecto de eliminar el ruido de los propios juicios será siempre una mejora de la exactitud predictiva.[17] Supongamos, por ejemplo, que la correlación entre los pronósticos que uno hace y un resultado es de .50 (PC = 67 %), pero el 50 por ciento de la varianza de los juicios consiste en ruido. Si pudiera hacer sus juicios sin ruido —como lo haría un modelo de él mismo—, su correlación con el mismo resultado se dispararía a .71 (PC =

75%). La reducción del ruido aumenta mecánicamente la validez del juicio predictivo.

En suma, la sustitución de uno mismo por un modelo de sí mismo hace dos cosas: elimina la sutileza y el ruido de patrón individual. El sólido hallazgo de que el modelo del juzgador es más válido que el propio juzgador transmite un importante mensaje: las ganancias de las reglas sutiles en el juicio humano —cuando existen— no suelen ser suficientes para compensar los efectos perjudiciales del ruido. Uno puede creer que él es más sutil y perspicaz, que repara más en los matices que la caricatura lineal de su pensamiento, pero la verdad es que la mayoría de las veces es más ruidoso.

¿Por qué las complejas reglas de predicción perjudican la exactitud, a pesar de que tenemos la penetrante sensación de que se basan en percepciones válidas? Por un lado, no es probable que muchas de las complejas reglas que inventa la gente sean en general ciertas. Sin embargo, hay otro problema: aun cuando las reglas complejas en principio sean válidas, se aplican de forma inevitable en condiciones que rara vez son observadas. Supongamos, por ejemplo, que hemos llegado a la conclusión de que merece la pena contratar a candidatos excepcionalmente originales, aunque sus puntuaciones en otros aspectos sean mediocres. El problema es que los candidatos excepcionalmente originales son, por definición, excepcionalmente raros. Como es probable que la evaluación de la originalidad no sea fiable, muchas puntuaciones altas en ese indicador son casualidades, y el talento del todo original suele pasar inadvertido. Las evaluaciones de rendimiento que podrían confirmar que los «originales» acaban siendo superestrellas también son imperfectas. Los errores de medición en ambos extremos disminuyen de forma inevitable la validez de las predicciones, y es muy probable que se pasen por alto sucesos imprevistos. Las ventajas de la verdadera sutileza se ahogan con rapidez en el error de medición.

Un estudio de Martin Yu y Nathan Kuncel presentó una versión más radical de la demostración de Goldberg.[18] Este estudio (que fue la base del ejemplo de Monica y Nathalie) utilizó datos de una firma consultora internacional que empleó a expertos para evaluar a ochocientos cuarenta y siete candidatos para puestos ejecutivos en tres muestras separadas. Los expertos puntuaron los resultados en siete facetas de evaluación distintas y

utilizaron su juicio clínico para asignar una puntuación predictiva global a cada una de ellas, con resultados muy poco impresionantes.

Yu y Kuncel decidieron no comparar a los juzgadores con el mejor modelo simple de sí mismos, sino con un modelo lineal «aleatorio». Generaron diez mil conjuntos de ponderaciones aleatorias para los siete predictores y aplicaron las diez mil fórmulas aleatorias para predecir el rendimiento profesional.[19]

Su sorprendente hallazgo fue que «cualquier» modelo lineal, cuando se aplicaba de forma coherente a todos los casos, era probable que superase a los juzgadores humanos en la predicción de un resultado a partir de la misma información. En una de las tres muestras, el 77 por ciento de los diez mil modelos lineales ponderados aleatoriamente lo hizo mejor que los expertos humanos. En las otras dos muestras, el 100 por ciento de los modelos aleatorios superó a los humanos. O, para decirlo sin rodeos, en ese estudio resultó casi imposible generar un modelo simple que lo hiciera peor que los expertos.

La conclusión de esta investigación es más sólida que la que extrajimos del trabajo de Goldberg sobre el modelo del juzgador, y eso que es un ejemplo extremo. En este marco, los juzgadores humanos obtuvieron resultados muy pobres en términos absolutos, lo que ayuda a explicar por qué incluso modelos lineales poco convincentes los superaron. Naturalmente, no debemos concluir que cualquier modelo gana a cualquier humano. Aun así, el hecho de que la adhesión maquinal a una regla simple (Yu y Kuncel la llaman «consistencia ciega») pueda mejorar de forma significativa el juicio en un problema difícil ilustra el enorme efecto del ruido en la validez de las predicciones clínicas.

Este rápido recorrido ha mostrado cómo el ruido perjudica al juicio clínico. En los juicios predictivos, los expertos humanos son fácilmente superados por fórmulas simples (modelos de la realidad, modelos de un juzgador y hasta modelos generados de manera aleatoria). Este descubrimiento aboga por el uso de métodos libres de ruido: reglas y algoritmos, que son el tema del siguiente capítulo.

- «La gente cree que capta la complejidad y añade sutileza cuando juzga. Sin embargo, la complejidad y la sutileza son en su mayor parte desperdicios, pues no aportan nada a la exactitud de los modelos simples.»
- «Más de sesenta años después de publicarse el libro de Paul E. Meehl, la idea de que la predicción mecánica es superior a las personas sigue siendo chocante.»
- «Hay tanto ruido en los juicios que un modelo sin ruido de un juzgador logra predicciones más exactas que el juzgador real.»

10
Reglas sin ruido

En los últimos años, la inteligencia artificial (IA) y, en particular, las técnicas de aprendizaje automático han permitido a las máquinas ejecutar muchas tareas que antes se consideraban humanas por antonomasia. Los algoritmos de aprendizaje automático pueden reconocer rostros, traducir idiomas e interpretar radiografías. Pueden resolver problemas de computación, como dar indicaciones instantáneas de conducción a miles de conductores con una rapidez y precisión asombrosas. Y, además, realizan complejas tareas de predicción: pronostican las decisiones del Tribunal Supremo de Estados Unidos, determinan qué acusados tienen más probabilidades de saltarse la libertad bajo fianza y evalúan qué llamadas a los servicios de protección infantil requieren con mayor urgencia la visita de un asistente social.

Aunque hoy estas son las aplicaciones que tenemos en mente cuando oímos la palabra «algoritmo», el término tiene un significado más amplio. Según la definición de un diccionario, un algoritmo es un «proceso o conjunto de reglas que se siguen en cálculos u otras operaciones de resolución de problemas, sobre todo los ejecutados por un ordenador». Según esta definición, los modelos simples y otras formas de juicio mecánico que describimos en el capítulo anterior también son algoritmos.

De hecho, son muchos los tipos de métodos mecánicos (desde reglas simples, casi irrisorias, hasta los algoritmos de máquinas más complejos e impenetrables) que pueden superar al juicio humano. Y una de las razones clave de esta capacidad —aunque no la única— es que todos los métodos mecánicos están libres de ruido.

Para examinar diferentes tipos de métodos basados en reglas y aprender cómo y en qué condiciones puede ser valioso cada método,

FIGURA 11. Cuatro tipos de reglas y algoritmos

MÁS SIMPLICIDAD: SÓLIDOS Y BELLOS

Robyn Dawes fue otro miembro del equipo de estrellas de Eugene (Oregón), que estudió los juicios en las décadas de 1960 y 1970. En 1974, Dawes logró un gran avance en la simplificación de las tareas de predicción. Su idea era sorprendente, casi herética: en vez de utilizar la regresión múltiple para determinar el peso preciso de cada predictor, propuso dar a todos los predictores el mismo peso.

Dawes etiquetó la fórmula del peso igual como «modelo lineal impropio». Su inesperado descubrimiento fue que esos modelos de peso igual eran tan exactos como los modelos de regresión «propios» y muy superiores a los juicios clínicos.[1]

Incluso los defensores de los modelos impropios admiten que esta afirmación es inverosímil y «contraria a la intuición estadística».[2] De hecho, Dawes y su ayudante Bernard Corrigan tuvieron dificultades inicialmente para publicar su artículo en las revistas científicas; los editores no les creyeron. Quien piense en el ejemplo de Monica y Nathalie del capítulo anterior, probablemente crea que algunos predictores son más importantes que otros. La mayoría de la gente daría, por ejemplo, más peso al liderazgo que a las habilidades técnicas. ¿Cómo puede una media directa no ponderada predecir el rendimiento de alguien mejor que una media cuidadosamente ponderada, o mejor que el juicio de un experto?

Hoy, muchos años después del avance de Dawes, se comprende bien el fenómeno estadístico que tanto sorprendió a sus contemporáneos. Como se ha explicado antes en este libro, la regresión múltiple calcula las ponderaciones «óptimas» que minimizan los errores al cuadrado, pero la regresión múltiple minimiza el error «en los datos originales». Por lo tanto, la fórmula se ajusta a sí misma para predecir cualquier aleatoriedad en los

datos. Si, por ejemplo, la muestra incluye a unos cuantos directivos que tienen altas capacidades técnicas y que también tuvieron un rendimiento excepcionalmente bueno por razones no relacionadas, el modelo exagerará el peso de la capacidad técnica.

El problema es que, cuando la fórmula se aplica «fuera de la muestra» —es decir, cuando se utiliza para predecir resultados en un conjunto de datos diferente—, las ponderaciones ya no serán óptimas. Las casualidades de la muestra original ya no están presentes, justamente porque eran casualidades; en la nueva muestra, los directivos con altas capacidades técnicas no son todos superestrellas. Y la nueva muestra tiene diferentes casualidades que la fórmula no puede predecir. La medida correcta de la exactitud predictiva de un modelo es su comportamiento en una nueva muestra, denominada su «correlación de validación cruzada». En efecto, un modelo de regresión tiene «demasiado» éxito en la muestra original, y una correlación de validación cruzada es casi siempre más baja de la que era en los datos originales. Dawes y Corrigan compararon los modelos de peso igual con los modelos de regresión múltiple (con validación cruzada) en varias situaciones. Uno de sus ejemplos contenía predicciones de la calificación media del primer año de noventa estudiantes de posgrado de la especialidad de Psicología en la Universidad de Illinois, que utilizaban diez variables relacionadas con el éxito académico: las puntuaciones de las pruebas de aptitud, las notas en la universidad, distintas valoraciones de los compañeros (por ejemplo, la extraversión) y varias valoraciones de uno mismo (por ejemplo, la diligencia). El modelo de regresión múltiple estándar obtuvo una correlación de .69, que se redujo a .57 (PC = 69 %) en validación cruzada. La correlación del modelo de peso igual con la calificación media del primer año fue casi la misma: .60 (PC = 70 %). En muchos otros estudios se han obtenido resultados similares.[3]

La pérdida de exactitud en la validación cruzada es peor cuando la muestra original es pequeña, porque las casualidades pesan más en las muestras pequeñas. El problema que señaló Dawes es que las muestras utilizadas en la investigación en ciencias sociales suelen ser tan pequeñas que la ventaja de la llamada «ponderación óptima» desaparece. Como dijo de forma memorable el estadístico Howard Wainer en el subtítulo de un artículo académico sobre la estimación de los pesos adecuados, «It Don’t

Make No Nevermind» (quería decir que no tenía ninguna tendencia o preferencia).[4] O, en palabras de Dawes, «no necesitamos modelos más precisos que nuestras mediciones».[5] Los modelos de peso igual funcionan bien porque no son susceptibles de sufrir accidentes de muestreo.

La inmediata implicación del trabajo de Dawes merece ser ampliamente conocida: se pueden hacer predicciones estadísticas válidas sin datos previos sobre el resultado que se intenta predecir. Todo lo que se necesita es una colección de predictores que permitan confiar en que estén correlacionados con el resultado.

Supongamos que hay que hacer predicciones sobre el rendimiento de ejecutivos que han sido calificados en una serie de aspectos, como en el ejemplo del capítulo 9. Se confía en que estas puntuaciones miden cualidades importantes, pero no se tienen datos sobre la capacidad de cada puntuación para predecir el rendimiento. Tampoco es posible permitirse el lujo de esperar unos años para seguir el rendimiento de una amplia muestra de ejecutivos. No obstante, podrían tomarse las siete puntuaciones, hacer el trabajo estadístico necesario para ponderarlas igual y utilizar el resultado como predicción. ¿Cuál sería este modelo de peso igual? Su correlación con el resultado sería de .25 (PC = 58 %), muy superior a la de las predicciones clínicas (r = .15, PC = 55 %), y seguramente bastante similar a la de un modelo de regresión con validación cruzada.[6] Además, no requiere ningún dato que no se tenga ni cálculo alguno complicado.

Para usar la expresión de Dawes, que se ha convertido en un meme entre los estudiosos del juicio, hay una «belleza sólida» en el peso igual.[7] La frase final del influyente artículo que introdujo la idea era otro lacónico resumen: «Todo el truco consiste en decidir qué variables considerar y luego saber cómo sumar».[8]

AÚN MÁS SIMPLICIDAD

Otro estilo de simplificación es el de los «modelos frugales» o «reglas simples». Los modelos frugales son modelos de la realidad que parecen cálculos de servilleta ridículamente simplificados. Sin embargo, en algunos casos, pueden hacer predicciones asombrosamente buenas.

Estos modelos se basan en una característica de la regresión múltiple que a la mayoría de la gente le resulta sorprendente. Supongamos que se utilizan

dos predictores que predicen fuertemente el resultado —sus correlaciones con el resultado son de .60 (PC = 71 %) y .55 (PC = 69 %)—. Supongamos también que los dos predictores están correlacionados entre sí con una correlación de .50. ¿Cómo será la predicción cuando los dos predictores se combinan de forma óptima? La respuesta es bastante decepcionante. La correlación es de .67 (PC = 73 %), más alta que antes, pero no mucho más.

Este ejemplo ilustra una regla general: la combinación de dos o más predictores correlacionados apenas es más predictiva que el mejor de ellos por sí solo. Dado que, en la vida real, los predictores están casi siempre correlacionados entre sí, este hecho estadístico respalda el uso de enfoques frugales de predicción que utilicen un número reducido de predictores. Las reglas sencillas que pueden aplicarse con pocos o ningún cálculo han generado predicciones extraordinariamente exactas en algunos ámbitos comparadas con modelos que utilizan muchos más predictores.

Un equipo de investigadores publicó en 2020 los resultados de un intento a gran escala de aplicar un enfoque frugal a una serie de problemas de predicción, entre ellos el de la elección a la que se enfrentan los jueces cuando han de decidir si liberar o retener a los acusados en espera de juicio. [9] Esa decisión es una predicción implícita del comportamiento del acusado. Si equivocadamente se le deniega la fianza, esa persona permanecerá detenida de forma innecesaria, con un coste importante para el individuo y para la sociedad. Y si se le concede equivocadamente la libertad bajo fianza, esa persona puede huir antes del juicio o incluso cometer otro delito.

El modelo que los investigadores construyeron utiliza solamente dos datos que se sabe que en gran medida predicen la probabilidad de que un acusado se aproveche de la libertad bajo fianza: la edad del acusado (las personas mayores tienen menos riesgo de fuga) y el número de citas judiciales anteriores a las que no se presentó (las personas que en el pasado no han comparecido tienden a reincidir). El modelo traduce estos dos *inputs* a un número de puntos que pueden utilizarse como puntuación de riesgo. El cálculo del riesgo para un acusado no necesita un ordenador (de hecho, ni siquiera una calculadora).

Cuando se puso a prueba este modelo frugal con un conjunto de datos reales, funcionó tan bien como los modelos que utilizan un número mucho

mayor de variables. El modelo frugal predijo el riesgo de fuga mejor que casi todos los jueces humanos.

El mismo enfoque frugal, que utiliza hasta cinco características ponderadas con números enteros pequeños (entre -3 y +3), se aplicó a tareas tan variadas como determinar la gravedad de un tumor a partir de datos mamográficos, diagnosticar enfermedades cardiacas y predecir el riesgo crediticio. En todas estas tareas, la regla frugal funcionó tan bien como los modelos de regresión más complejos (aunque, en general, no tan bien como el aprendizaje automático).

En otra demostración del poder de las reglas simples, otro equipo de investigadores estudió un problema judicial algo distinto: la predicción de la reincidencia.[10] Utilizando solo dos *inputs*, fue capaz de igualar la validez de una herramienta existente que emplea ciento treinta y siete variables para evaluar el nivel de riesgo de un acusado.[11] No sorprende que estos dos predictores (la edad y el número de condenas previas) estén estrechamente relacionados con los dos factores utilizados en el modelo de la libertad bajo fianza, y su asociación con el comportamiento delictivo está muy documentada.[12]

El atractivo de las reglas frugales es que son transparentes y fáciles de aplicar. Además, esas ventajas se obtienen con un coste en exactitud relativamente bajo en comparación con modelos más complejos.

MÁS COMPLEJIDAD: HACIA EL APRENDIZAJE AUTOMÁTICO

En la segunda parte de nuestro viaje, vamos a ir ahora en la dirección opuesta en el espectro de la complejidad. ¿Y si pudiéramos utilizar muchos más predictores, reunir muchos más datos sobre cada uno de ellos, encontrar patrones de relación que ningún humano podría detectar y modelizar estos patrones para lograr una mejor predicción? Esta es, en esencia, la promesa de la inteligencia artificial (IA).

Los conjuntos de datos muy grandes son esenciales para realizar análisis complejos, y la creciente disponibilidad de dichos conjuntos de datos es una de las principales causas del rápido progreso de la IA en los últimos años.[13] Por ejemplo, los grandes conjuntos de datos permiten tratar mecánicamente con «excepciones de pierna rota». Esta expresión, un tanto críptica, se remonta a un ejemplo que Meehl imaginó: consideremos un

modelo diseñado para predecir la probabilidad de que la gente vaya al cine una noche. Con independencia de la confianza que se tenga en el modelo, si se sabe que una persona concreta acaba de romperse una pierna, probablemente sepa mejor que el modelo cómo será su noche.

Cuando se utilizan modelos simples, el principio de la pierna rota encierra una importante lección para los responsables de la toma de decisiones: les indica cuándo deben desechar el modelo y cuándo no. Si tienen información decisiva que el modelo no ha podido tener en consideración, se trata de una verdadera pierna rota y deben desechar la recomendación del modelo. Por otra parte, a veces no estarán de acuerdo con la recomendación de un modelo, aunque carezcan de esa información privada. En estos casos, su tentación de desechar el modelo refleja un patrón personal que están aplicando a los mismos predictores. Como es muy probable que este patrón personal no sea válido, deberían abstenerse de desechar el modelo; es probable que su intervención empeore la predicción.

Una de las razones del éxito de los modelos de aprendizaje automático en tareas de predicción es que son capaces de descubrir esas piernas rotas, muchas más de las que los humanos pueden imaginar. Dada una enorme cantidad de datos sobre un gran número de casos, un modelo que rastree el comportamiento de los cinéfilos podría realmente aprender, por ejemplo, que las personas que han acudido al hospital en su día habitual de cine es poco probable que vean una película esa noche. Esta manera de mejorar las predicciones de episodios infrecuentes reduce la necesidad de supervisión humana.

Este proceso no es mágico, ni siquiera lo es de comprensión, sino de mera búsqueda de patrones. Aunque debemos admirar el poder del aprendizaje automático, debemos recordar que probablemente una IA se tomará algún tiempo para entender «por qué» una persona que se ha roto una pierna se perderá su noche de cine.

UN EJEMPLO: LAS MEJORES DECISIONES SOBRE LIBERTAD BAJO FIANZA

Casi al mismo tiempo que el equipo de investigadores antes mencionado aplicaba reglas simples al problema de las decisiones sobre la libertad bajo fianza, otro equipo, dirigido por Sendhil Mullainathan, entrenaba complejos

modelos de IA para la misma tarea.[14] El equipo de IA tuvo acceso a un conjunto mayor de datos (758.027 decisiones sobre libertad bajo fianza). En cada uno de los casos, el equipo pudo acceder a la información de la que también disponía el juez: el delito del acusado, sus antecedentes y sus incomparecencias anteriores. Salvo la edad, no se utilizó ninguna otra información demográfica para entrenar el algoritmo. Los investigadores también supieron, para cada uno de los casos, si el acusado fue puesto en libertad y, en caso afirmativo, si no compareció ante el tribunal o volvió a ser detenido. (De todos los acusados, el 74 por ciento quedó en libertad, y, de estos, el 15 por ciento no compareció ante el tribunal y el 26 por ciento volvió a ser detenido.) Con estos datos, los investigadores entrenaron un algoritmo de aprendizaje automático y evaluaron su capacidad.[15] Como el modelo se construyó mediante aprendizaje automático, no se limitó a las combinaciones lineales. Si detectaba una regularidad más compleja en los datos, podía utilizar ese patrón para mejorar sus predicciones.

El modelo fue diseñado para generar una predicción del riesgo de fuga cuantificado en una puntuación numérica, en lugar de una decisión de fianza-no fianza. Este enfoque reconoce que el umbral de riesgo máximo aceptable, es decir, el nivel de riesgo por encima del cual se debe denegar la fianza a un acusado, requiere un juicio evaluativo que un modelo no puede realizar. Sin embargo, los investigadores calcularon que, con independencia de dónde se fije el umbral de riesgo, el uso de la puntuación predictiva de su modelo supondría una mejora con respecto a la actuación de los jueces humanos. Si el umbral de riesgo se fija de forma que el número de personas a las que se deniega la fianza siga siendo el mismo que cuando deciden los jueces —calculó el equipo de Mullainathan—, los índices de delincuencia podrían reducirse hasta en un 24 por ciento, porque las personas entre rejas serían las más propensas a reincidir. Si, por el contrario, el umbral de riesgo se fija para reducir al máximo el número de personas a las que se deniega la fianza sin que aumente la delincuencia, los investigadores calcularon que el número de personas detenidas podría reducirse hasta en un 42 por ciento. En otras palabras, el modelo de aprendizaje automático funciona mucho mejor que los jueces humanos a la hora de predecir qué acusados son de mayor riesgo.

El modelo construido empleando el aprendizaje automático tuvo también mucho más éxito que los modelos lineales que utilizaban la misma

información. La razón es interesante: «El algoritmo de aprendizaje automático encuentra señales significativas en combinaciones de variables que sin él podrían pasarse por alto».[16] La capacidad del algoritmo para encontrar patrones que otros métodos pasan fácilmente por alto es muy pronunciada para los acusados que el algoritmo clasifica como de máximo riesgo. En otras palabras, algunos patrones en los datos, aunque raros, predicen claramente el riesgo alto. Este descubrimiento —que el algoritmo capta patrones raros, pero decisivos— nos devuelve a la idea de la pierna rota.

Los investigadores utilizaron también el algoritmo para construir un modelo de cada juez análogo al modelo del juez que describimos en el capítulo 9 (pero sin limitarse a las combinaciones lineales simples). La aplicación de estos modelos a todo el conjunto de datos permitió al equipo simular las decisiones que habrían tomado los jueces si hubieran visto los mismos casos y compararlas. Los resultados indicaron un considerable ruido del sistema en las decisiones sobre la fianza. Una parte era ruido de nivel: cuando los jueces se clasifican por indulgencia, el quintil más indulgente (es decir, el 20 por ciento de los jueces que tienen las tasas más altas de dictar libertad) liberó al 83 por ciento de los acusados, mientras que el quintil menos indulgente liberó solo al 61 por ciento. Los jueces también tienen patrones de juicio muy diferentes con respecto a qué acusados presentan mayor riesgo de fuga. Un acusado al que un juez considera con bajo riesgo de fuga puede ser considerado de alto riesgo de fuga por otro juez que en general no es más estricto. Estos resultados ofrecen una clara evidencia del ruido de patrón. Un análisis más detallado reveló que las diferencias entre los casos representaban el 67 por ciento de la varianza, y el ruido del sistema, el 33 por ciento. El ruido del sistema incluía algo de ruido de nivel, es decir, diferencias de severidad media, pero la mayor parte de él (el 79 por ciento) era ruido de patrón.[17]

Por último, y por fortuna, la mayor exactitud del programa de aprendizaje automático no se consigue a costa de otros objetivos identificables que podrían haber perseguido los jueces (en particular, la equidad racial). En teoría, aunque el algoritmo no utiliza datos raciales, el programa podría agravar inadvertidamente las disparidades raciales. Estas últimas podrían surgir si el modelo utilizara predictores muy correlacionados con la raza (como el código postal) o si la fuente de los datos con los que se entrena el

algoritmo está sesgada. Si, por ejemplo, se utiliza el número de arrestos anteriores como predictor, y si los arrestos anteriores se ven afectados por la discriminación racial, el algoritmo resultante también discriminará.

Aunque, en principio, este tipo de discriminación es sin duda un riesgo, en algunos aspectos importantes las decisiones de este algoritmo tienen menos sesgo racial que las de los jueces, no más. Por ejemplo, si el umbral de riesgo se fija para alcanzar la misma tasa de delincuencia que las decisiones de los jueces, el algoritmo encarcela a un 41 por ciento menos de personas de color. Los resultados son similares en otros escenarios: la ganancia de exactitud no tiene por qué exacerbar las disparidades raciales y, como demostró también el equipo de investigación, al algoritmo se le puede enseñar fácilmente para reducirlas.

Otro estudio en un ámbito diferente ilustra cómo los algoritmos pueden aumentar la exactitud y reducir la discriminación de forma simultánea. Bo Cowgill, profesor de la Columbia Business School, estudió la contratación de ingenieros de *software* en una gran empresa tecnológica.[18] En lugar de utilizar examinadores de currículos (humanos) para seleccionar a quiénes se haría una entrevista, Cowgill desarrolló un algoritmo de aprendizaje automático para examinar los currículos de los candidatos, y lo entrenó con más de trescientos mil que la empresa había recibido y evaluado. Los candidatos seleccionados por el algoritmo tuvieron un 14 por ciento más de probabilidades de recibir una oferta de trabajo tras las entrevistas que los seleccionados por humanos. Cuando los candidatos recibían ofertas, en el grupo del algoritmo había un 18 por ciento más de probabilidades de aceptarlas que en el grupo seleccionado por humanos. El algoritmo también eligió un grupo de candidatos más variado en términos de raza, sexo y otras métricas; era mucho más probable que seleccionara candidatos «no tradicionales», como los que no se graduaron en una escuela de élite, los que carecían de experiencia laboral previa y los que no tenían recomendaciones. Los seres humanos tienden a preferir los currículos que marcan todas las casillas del perfil «típico» de un ingeniero de *software*, pero el algoritmo dio a cada predictor relevante su peso adecuado.

Para ser claros, estos ejemplos no demuestran que los algoritmos sean siempre equitativos, no sesgados o no discriminatorios. Un ejemplo conocido es un algoritmo que se supone que predice el éxito de los candidatos a un puesto de trabajo, pero que en realidad se entrena con una

muestra de decisiones de promoción anteriores. Por supuesto, un algoritmo de este tipo reproducirá todos los sesgos humanos en las decisiones de promoción del pasado.

Es posible, y quizá demasiado fácil, construir un algoritmo que perpetúe las disparidades raciales o sexuales, y se han registrado muchos casos de algoritmos que hicieron justo eso. La visibilidad de estos casos explica la creciente preocupación por los sesgos en la toma de decisiones basadas en algoritmos. Sin embargo, antes de sacar conclusiones generales sobre estos, debemos recordar que algunos de ellos no solo son más exactos que los jueces humanos, sino también más justos.

¿POR QUÉ NO USAMOS REGLAS MÁS A MENUDO?

Para resumir este breve recorrido por la toma de decisiones mecánica, volvamos sobre dos razones de la superioridad de las reglas de todo tipo sobre el juicio humano. En primer lugar, como hemos descrito en el capítulo 9, todas las técnicas de predicción mecánica, no solo las más recientes y complejas, representan mejoras significativas con respecto al juicio humano. La combinación de patrones personales y ruido de ocasión pesa tanto en la calidad del juicio humano que la simplicidad y la ausencia de ruido son ventajas considerables. Las reglas sencillas que son solamente sensatas dan por regla general mejores resultados que el juicio humano.

En segundo lugar, los datos son a veces lo bastante ricos como para que las complejas técnicas de IA detecten patrones válidos y vayan mucho más allá del poder predictivo de un modelo simple. Cuando la IA tiene éxito en este sentido, la ventaja de estos modelos sobre el juicio humano no es solo la ausencia de ruido, sino también la capacidad de explotar mucha más información.

Dadas estas ventajas y la enorme cantidad de pruebas que las respaldan, cabe preguntarse por qué los algoritmos no se utilizan con mucha más frecuencia para los tipos de juicios profesionales de los que tratamos en este libro. Por muy animadas que sean las discusiones sobre los algoritmos y el aprendizaje automático, y a pesar de importantes excepciones en determinados campos, su uso sigue siendo limitado. Muchos expertos ignoran el debate sobre las diferencias entre el juicio clínico y el mecánico, y prefieren confiar en su juicio particular. Tienen fe en sus intuiciones y

dudan de que las máquinas puedan hacerlo mejor. Consideran que la idea de la toma de decisiones mediante algoritmos es deshumanizadora e implica una renuncia a su responsabilidad.

El uso de algoritmos en el diagnóstico médico, por ejemplo, aún no es rutinario, a pesar de los considerables avances, y pocas organizaciones utilizan algoritmos en sus decisiones de contratación y promoción. Los ejecutivos de los estudios de Hollywood dan luz verde a las películas basándose en su juicio y su experiencia, no en una fórmula. Los editores de libros hacen lo propio. Y, si la historia del equipo de béisbol de los Oakland Athletics, obsesionado con las estadísticas, contada en el superventas de Michael Lewis *Moneyball*, ha causado tanto impacto, es justo porque el rigor algorítmico había sido durante mucho tiempo la excepción, no la regla, en el proceso de toma de decisiones de los equipos deportivos. Incluso hoy, entrenadores, directivos y personas que trabajan con ellos suelen confiar en su instinto e insisten en que el análisis estadístico no puede sustituir al sentido común.

En un artículo de 1996, Meehl y un coautor enumeraron (y rebatieron) no menos de diecisiete tipos de objeciones que psiquiatras, médicos, jueces y otros profesionales ponían al juicio mecánico.[19] Los autores concluyeron que la resistencia de los enjuiciadores clínicos podía explicarse por una combinación de factores sociopsicológicos, entre ellos su «temor al desempleo tecnológico», su «escasa formación» y una «aversión general a los ordenadores».

Desde entonces, los investigadores han identificado otros factores que contribuyen a esta resistencia. No pretendemos ofrecer aquí una relación completa de esta investigación. Nuestro objetivo en este libro es ofrecer sugerencias para mejorar el juicio humano, no defender la «sustitución de las personas por las máquinas», como habría dicho el juez Frankel.

Sin embargo, algunos descubrimientos sobre el origen de la resistencia humana a la predicción mecánica son relevantes para nuestro debate sobre el juicio humano. Una idea clave ha surgido de investigaciones recientes: la gente no desconfía sistemáticamente de los algoritmos. Cuando se le da a elegir entre el consejo de un humano y el de un algoritmo, por ejemplo, suele preferir el algoritmo.[20] La resistencia a los algoritmos, o «aversión a los algoritmos», no siempre se manifiesta en un rechazo generalizado a la adopción de nuevas herramientas de apoyo a la toma de decisiones. Más a

menudo, la gente está dispuesta a dar una oportunidad a un algoritmo, pero deja de confiar en él en cuanto ve que comete errores.[21]

En cierta manera, esta reacción parece sensata: ¿por qué molestarse con un algoritmo en el que no se puede confiar? Como humanos, somos muy conscientes de que cometemos errores, pero ese es un privilegio que no estamos dispuestos a compartir. Esperamos que las máquinas sean perfectas. Si esta expectativa queda defraudada, las descartamos.[22]

Sin embargo, debido a esta expectativa intuitiva, es probable que las personas desconfíen de los algoritmos y que sigan utilizando su criterio, incluso cuando esta elección produce resultados que se han demostrado que son inferiores. Esta actitud está muy arraigada, y es poco probable que cambie hasta que no se consiga una precisión predictiva casi perfecta.

Por fortuna, gran parte de lo que hace que las reglas y los algoritmos sean mejores puede reproducirse en el juicio humano. No podemos esperar utilizar la información de forma tan eficiente como lo hace un modelo de IA, pero podemos esforzarnos por emular la simplicidad y la ausencia de ruido de los modelos simples. En la medida en que podamos adoptar métodos que reduzcan el ruido del sistema, veríamos mejoras en la calidad de los juicios predictivos. Cómo mejorar nuestros juicios es el tema principal de la quinta parte de este libro.

SOBRE REGLAS Y ALGORITMOS

- «Cuando hay muchos datos, los algoritmos de aprendizaje automático funcionarán mejor que los humanos y mejor que los modelos simples. Sin embargo, incluso las reglas y los algoritmos más sencillos tienen grandes ventajas sobre los juzgadores humanos: están libres de ruido y no intentan reflejar ideas complejas, normalmente sin validez, sobre los predictores.»
- «Dado que carecemos de datos sobre el resultado que debemos predecir, ¿por qué no utilizamos un modelo de peso igual? Puede que no lo haga tan bien como un modelo adecuado, pero seguirá siendo mejor que el juicio humano caso por caso.»
- «No estás de acuerdo con la predicción del modelo, lo entiendo. Pero ¿hay aquí una pierna rota, o simplemente no te gusta la predicción?»

- «El algoritmo comete errores, por supuesto. Sin embargo, si los juzgadores humanos cometen errores aún mayores, ¿en quién debemos confiar?»

11
Ignorancia objetiva

A menudo hemos tenido la experiencia de compartir con un público de ejecutivos el material de los dos últimos capítulos con sus aleccionadoras conclusiones sobre los limitados logros del juicio humano. El mensaje que nos proponemos transmitir existe desde hace más de medio siglo, y sospechamos que pocos responsables de decisiones han evitado exponerse a él. Sin embargo, ocurre que son capaces de resistirlo.

Algunos de los ejecutivos de nuestro público nos dicen con orgullo que confían más en su instinto que en cualquier serie de análisis. Muchos otros son menos contundentes, pero comparten la misma opinión. La investigación sobre la toma de decisiones en la gestión empresarial ha demostrado que los ejecutivos, en particular los más veteranos y experimentados, recurren en gran medida a lo que llaman «intuición», «corazonada» o, simplemente, «juicio» (en un sentido diferente del que el vocablo tiene en este libro).[1]

A los responsables de la toma de decisiones les gusta, pues, escuchar a su instinto, y en su mayoría parecen estar contentos con lo que oyen. Lo cual suscita una pregunta: ¿qué es exactamente lo que escuchan en su fuero interno estas personas bendecidas con la combinación de autoridad y gran confianza en sí mismas?

Una crítica de la intuición en la toma de decisiones en la gestión empresarial la define como «un juicio para un determinado curso de acción que viene a la mente con un aura o convicción de corrección o plausibilidad, pero sin razones o justificaciones claramente articuladas; exactamente un “saber”, pero sin saber por qué».[2] Proponemos que esta sensación de saber sin saber por qué no es sino la «señal interna» de consumación del juicio que mencionamos en el capítulo 4.

La señal interna es una recompensa autoadministrada que las personas se esfuerzan (o a veces no tanto) por conseguir cuando han de emitir un juicio. Es una experiencia emocional que las satisface, una agradable sensación de coherencia porque creen que las pruebas consideradas y el juicio alcanzado

son correctos. Todas las piezas del puzle parecen encajar. (Más adelante veremos que esta sensación de coherencia se refuerza a menudo ocultando o ignorando las piezas que no encajan.)

Lo que hace que la señal interna sea importante —y engañosa— es que no se interpreta como una sensación, sino como una creencia. Esta experiencia emocional («la evidencia sienta bien») se disfraza de confianza racional en la validez del propio juicio («lo sé, aunque no sepa por qué»).

Sin embargo, la confianza no es garantía de exactitud, y muchas predicciones confiadas resultan erróneas. Aunque tanto el sesgo como el ruido contribuyen a los errores de predicción, la mayor fuente de estos errores no es que la calidad de los juicios predictivos «sea» limitada, sino lo limitada que «puede ser». Esta limitación, que llamamos «ignorancia objetiva», es el centro del siguiente capítulo.

IGNORANCIA OBJETIVA

Hay una cuestión que podemos plantearnos cuando hacemos juicios predictivos repetidos. Esta cuestión es aplicable a cualquier tarea: elegir acciones en bolsa, por ejemplo, o predecir el rendimiento de atletas profesionales. Para simplificar, elegiremos el mismo ejemplo que utilizamos en el capítulo 9: la selección de candidatos a un puesto en una empresa. Imaginemos que hemos evaluado a un centenar de candidatos a lo largo de los años. Ahora tenemos la oportunidad de evaluar lo acertadas que fueron nuestras decisiones en comparación con las evaluaciones que hicimos del rendimiento, evaluado objetivamente, de los candidatos desde entonces. Si elegimos un par de candidatos al azar, ¿con qué frecuencia coincidirían nuestro juicio *ex ante* y las evaluaciones *ex post*? En otras palabras: al comparar dos candidatos cualesquiera, ¿cuál es la probabilidad de que el que creíamos que poseía más potencial resulte ser de hecho el de mayor rendimiento?

A menudo encuestamos de manera informal a grupos de ejecutivos sobre esta cuestión. Las respuestas más frecuentes se sitúan en la franja del 75-85 por ciento, y sospechamos que estas respuestas se hallan limitadas por la modestia y por el deseo de no parecer jactancioso. Las conversaciones privadas e individuales indican que la verdadera sensación de confianza suele ser mayor.

Ahora que estamos familiarizados con la estadística del porcentaje de concordancia, podemos ver fácilmente el problema que plantea esta evaluación. Un PC del 80 por ciento corresponde, aproximadamente, a una correlación de .80. Este nivel de poder predictivo rara vez se alcanza en el mundo real. En el ámbito de la selección de personal, un estudio reciente ha revelado que la capacidad de los juzgadores humanos no se acerca a esa cifra. Por término medio, alcanzan una correlación predictiva de .28 (PC = 59 %).[3]

Si consideramos el reto que supone la selección de personal, los resultados decepcionantes no sorprenden tanto. Una persona que comienza un nuevo trabajo se encontrará con buena cantidad de escollos y de oportunidades, y el azar intervendrá para cambiar el rumbo de su vida de muchas maneras. Puede encontrarse con un supervisor que cree en ella, le da oportunidades, promueve su trabajo e impulsa su confianza en sí misma y su motivación. También puede tener menos suerte y, sin culpa alguna, comenzar su carrera con un fracaso desmoralizador. También en su vida personal puede haber acontecimientos que afecten a su rendimiento. Nadie puede predecir hoy ninguno de estos acontecimientos y circunstancias; ni el mejor modelo de predicción del mundo puede hacerlo. Esta irremediable incertidumbre hace que la suma total de lo que afectará al resultado que se está prediciendo simplemente no pueda conocerse de antemano.

Además, hay muchas cosas sobre los candidatos que en un principio se pueden conocer, pero que se desconocen en el momento de emitir un juicio sobre ellos. Para nuestros fines, no importa si estas lagunas de conocimiento provienen de la falta de test suficientemente predictivos, de la decisión de que el coste de adquirir más información no está justificado o de alguna negligencia en la búsqueda de datos. De un modo u otro, el grado de información no es el ideal.

Tanto la incertidumbre irremediable (lo que no es posible saber) como la información imperfecta (lo que se podría saber, pero no se sabe) impiden la predicción perfecta. Estos desconocimientos no son problemas de sesgo o ruido en el juicio sobre los candidatos; son características objetivas de la tarea. Esta ignorancia objetiva de datos importantes limita en gran medida la exactitud alcanzable. Aquí nos tomamos la libertad de sustituir el término comúnmente usado de «incertidumbre» por el de «ignorancia». Este término ayuda a limitar el riesgo de confusión entre la incertidumbre, que se

refiere al mundo y al futuro, y el ruido, que es la variabilidad en juicios que deberían ser idénticos.

Hay más información (y menos ignorancia objetiva) en unas situaciones que en otras. La mayoría de los juicios profesionales son bastante buenos. Para muchas enfermedades, las predicciones de los médicos son excelentes, y para muchos conflictos legales, los abogados pueden decir, con gran exactitud, cuál será el fallo de los jueces.

Sin embargo, en general, se puede esperar con seguridad que las personas que se dedican a tareas de predicción subestimen su ignorancia objetiva. El exceso de confianza es uno de los sesgos cognitivos mejor documentados. En particular, los juicios sobre la propia capacidad de hacer predicciones precisas, aun a partir de información limitada, son excesivamente confiados.[4] Lo que dijimos del ruido en los juicios predictivos puede también decirse de la ignorancia objetiva: dondequiera que haya predicción, hay ignorancia, y más de la que se piensa.

EXPERTOS DEMASIADO CONFIADOS

Un buen amigo nuestro, el psicólogo Philip Tetlock, ha estado siempre provisto de un inquebrantable compromiso con la verdad y de un malicioso sentido del humor. En 2005 publicó un libro titulado *Expert Political Judgment*. A pesar de su título, que parece neutral, el libro era un devastador ataque a las capacidades de los expertos para hacer predicciones acertadas sobre los acontecimientos políticos.

Tetlock estudió las predicciones de casi trescientos expertos: periodistas reputados, académicos respetados y asesores de alto nivel de líderes nacionales. Preguntó si sus predicciones políticas, económicas y sociales se cumplían. La investigación le llevó dos décadas: para saber si las predicciones a largo plazo son correctas, se necesita paciencia.

El hallazgo clave de Tetlock fue que, en sus predicciones sobre los principales acontecimientos políticos, los supuestos expertos son poco convincentes, y en un grado asombroso. El libro se hizo célebre por su sorprendente remate: «El experto medio acertaba tanto como un chimpancé lanzando dardos». Una enunciación más precisa del mensaje del libro era que los expertos que se ganan la vida «comentando u ofreciendo consejos sobre tendencias políticas y económicas» no eran «mejores que los

periodistas o los lectores atentos de *The New York Times* en la "interpretación" de nuevas situaciones».[5] Sin duda, los expertos contaban grandes historias. Podían analizar una situación, pintar un cuadro convincente de cómo evolucionaría y refutar, con gran confianza, en los estudios de televisión, las objeciones de aquellos que no estaban de acuerdo con ellos. Pero ¿sabían realmente lo que pasaría? Apenas.

Tetlock llegó a esta conclusión abriéndose camino entre las historias existentes. En cada tema, pidió a los expertos que asignaran probabilidades a tres posibles resultados: *statu quo*, más de algo o menos de eso mismo. Un chimpancé lanzador de dardos «elegiría» cada uno de estos resultados con la misma probabilidad —un tercio—, con independencia de la realidad. Los expertos de Tetlock apenas superaron este estándar tan bajo. En promedio, asignaron probabilidades un poco más altas a los acontecimientos que ocurrieron que a los que no ocurrieron, pero la característica más destacada de su tarea era una excesiva confianza en sus predicciones. Los expertos bendecidos con teorías claras sobre el funcionamiento del mundo eran los más confiados y los menos precisos.

Los hallazgos de Tetlock indican que las predicciones detalladas a largo plazo sobre acontecimientos concretos son sencillamente imposibles. El mundo es un lugar desordenado donde acontecimientos menores pueden tener grandes consecuencias. Por ejemplo, considérese el hecho de que, en el instante de la concepción, hubo una probabilidad de que todos los personajes significativos de la historia (y también los insignificantes) nacieran con un sexo diferente. Los acontecimientos imprevisibles están destinados a ocurrir, y las consecuencias de estos acontecimientos imprevisibles también son imprevisibles. En consecuencia, la ignorancia objetiva se acumula continuamente cuanto más se mira hacia el futuro. El límite del juicio político de los expertos no lo pone la limitación cognitiva de los analistas, sino su irremediable ignorancia objetiva del futuro.

Nuestra conclusión es, por tanto, que no hay que culpar a los expertos de los fracasos de sus predicciones a largo plazo. Sin embargo, merecen algunas críticas por intentar una tarea imposible y creer que pueden tener éxito en ella.

Algunos años después de su sorprendente descubrimiento de la inutilidad de muchos pronósticos a largo plazo, Tetlock formó equipo con su esposa, Barbara Mellers, para estudiar la capacidad de acierto de las personas

cuando se les pide que pronostiquen los acontecimientos mundiales para un plazo relativamente corto, normalmente inferior a un año. El equipo descubrió que el pronóstico a corto plazo es difícil, pero no imposible, y que algunas personas, a las que Tetlock y Mellers llamaron «superanalistas», son sistemáticamente mejores que la mayoría de las demás, incluidos los profesionales de los servicios de inteligencia. En los términos que utilizamos aquí, sus nuevos hallazgos eran compatibles con la noción de que la ignorancia objetiva aumenta a medida que miramos al futuro. Volveremos a hablar de los superanalistas en el capítulo 21.

JUZGADORES POBRES Y MODELOS POCO MEJORES

Las primeras investigaciones de Tetlock demostraron la incapacidad general de las personas para hacer predicciones políticas a largo plazo. Hasta una persona con una bola de cristal habría cambiado sus conclusiones por completo. Una tarea solo puede considerarse imposible después de que muchos actores creíbles la hayan intentado y hayan fracasado. Si, como hemos demostrado, la suma mecánica de información es a menudo superior al juicio humano, la exactitud predictiva de reglas y algoritmos proporciona una prueba mejor de lo intrínsecamente predecibles o impredecibles que pueden ser los resultados.

Los capítulos precedentes pueden haber dejado la impresión de que los algoritmos son, de forma abrumadora, superiores a los juicios predictivos, pero esta impresión sería engañosa. Las modelos son sistemáticamente mejores que las personas, pero no mucho más. No hay en lo fundamental pruebas de situaciones en las que, con idéntica información, las personas lo hagan muy mal y los modelos lo hagan muy bien.

En el capítulo 9, mencionamos una revisión de ciento treinta y seis estudios que demostraban la superioridad de la suma mecánica sobre el juicio clínico.[6] Aunque las pruebas de esa superioridad son, de hecho, «muy numerosas y consistentes», la diferencia de capacidad no es grande. Noventa y tres de los estudios de dicha revisión se centraron en decisiones binarias y midieron la «tasa de aciertos» de los juicios clínicos y las fórmulas. En el estudio de la mediana, los juicios clínicos acertaron el 68 por ciento de las veces y las fórmulas el 73 por ciento. Un subconjunto más pequeño de treinta y cinco estudios utilizó el coeficiente de correlación

como medida de acierto. En estos estudios, los juicios clínicos alcanzaron una correlación media con el resultado de .32 (PC = 60 %), mientras que la de las fórmulas fue de .56 (PC = 69 %). En ambas métricas, las fórmulas son sistemáticamente mejores que los juicios clínicos, pero sigue llamando la atención la limitada validez de las predicciones mecánicas. La capacidad de los modelos no cambia el panorama de un techo de predictibilidad bastante bajo.

¿Y qué decir de la inteligencia artificial (IA)? Como hemos señalado, la IA obtiene a menudo mejores resultados que los modelos más sencillos. Sin embargo, en la mayoría de las aplicaciones, su aptitud dista mucho de ser perfecta. Consideremos, por ejemplo, el algoritmo de predicción de la libertad bajo fianza que discutimos en el capítulo 10. Allí observamos que, manteniendo constante el número de personas a las que se les denegó la fianza, el algoritmo podía reducir los índices de delincuencia hasta en un 24 por ciento. Se trata de una impresionante mejora con respecto a las predicciones de los jueces humanos, pero si el algoritmo pudiera predecir con una exactitud perfecta qué acusados reincidirán, podría reducir mucho más la tasa de delincuencia. Las predicciones de futuros delitos en la película *Minority Report* son ciencia ficción por una razón: hay una gran cantidad de ignorancia objetiva en la predicción del comportamiento humano.

Otro estudio, dirigido por Sendhil Mullainathan y Ziad Obermeyer, hizo un modelo del diagnóstico de infartos.[7] Cuando los pacientes presentan signos de un posible infarto, los médicos de urgencias deben decidir si prescriben más pruebas. En principio, los pacientes deben someterse a las pruebas solo cuando los riesgos del infarto son lo bastante elevados: como las pruebas no solo son costosas, sino también invasivas y arriesgadas, no son deseables para los pacientes de bajo riesgo. Por tanto, la decisión del médico de hacer esas pruebas requiere una evaluación de los riesgos de infarto. Los investigadores construyeron un modelo de IA para llevar a cabo esta evaluación. El modelo utiliza más de dos mil cuatrocientas variables y se basa en una amplia muestra de casos (4,4 millones de visitas a Medicare por 1,6 millones de pacientes). Con tal cantidad de datos, es probable que el modelo se acerque a los límites de la ignorancia objetiva.

No es sorprendente que el modelo de IA sea claramente superior al de los médicos. Para evaluar la capacidad del modelo, consideremos a los

pacientes a los que el modelo situó en el decil más alto de riesgo. Cuando se hicieron pruebas a estos pacientes, resultó que el 30 por ciento de ellos había sufrido un infarto, mientras que los pacientes situados en el centro de la distribución de riesgo que lo habían sufrido fueron el 9,3 por ciento. Este nivel de discriminación es sorprendente, pero también está lejos de ser perfecto. Podemos concluir de manera razonable que la actuación de los médicos está limitada por las restricciones de la ignorancia objetiva tanto al menos como por los defectos de sus juicios.

LA NEGACIÓN DE LA IGNORANCIA

Al insistir en la imposibilidad de una predicción perfecta, podría parecer que estamos diciendo lo obvio. Es cierto que afirmar que el futuro es imprevisible no es un avance conceptual. Sin embargo, la obviedad de este hecho solo se ve igualada por la regularidad con la que se ignora, como demuestran los continuos hallazgos de exceso de confianza en la predicción.

La frecuencia del exceso de confianza arroja una nueva luz sobre nuestra encuesta informal a los responsables de la toma de decisiones que confían en el instinto. Hemos observado que la gente suele confundir su sensación subjetiva de confianza con una indicación de validez predictiva. Después de exponer en el capítulo 9 las pruebas sobre Nathalie y Monica, por ejemplo, la señal interna que el lector percibió cuando llegó a un juicio coherente le hizo confiar en que Nathalie era la candidata más indicada. Sin embargo, confiando en esa predicción, cayó en la ilusión de validez: la exactitud que pudo alcanzar con la información que le dimos era bastante baja.

Las personas que se creen capaces de un nivel tremendamente alto de exactitud predictiva no solo tienen demasiada confianza en sí mismas; no niegan el riesgo de ruido y el sesgo en sus juicios y tampoco se consideran simplemente superiores a los demás mortales. Sin embargo, creen en la predictibilidad de acontecimientos que, de hecho, son impredecibles y niegan implícitamente la realidad de la incertidumbre. En los términos que aquí hemos utilizado, esta actitud equivale a una «negación de la ignorancia».

La negación de la ignorancia añade una respuesta al rompecabezas que desconcertaba a Meehl y a sus seguidores: por qué su mensaje ha

permanecido en gran medida ignorado y por qué los responsables de la toma de decisiones siguen confiando en su intuición. Cuando estos escuchan a su instinto, oyen la señal interna y sienten la recompensa emocional que les proporciona. Esta señal interna de que se ha llegado a un buen juicio es la voz de la confianza, del «saber sin saber por qué». Pero una evaluación objetiva del verdadero poder predictivo de las evidencias rara vez justifica ese nivel de confianza.

No es fácil renunciar a la recompensa emocional de la certeza intuitiva. Es revelador que los líderes afirmen que son especialmente propensos a recurrir a las decisiones intuitivas en situaciones que perciben como muy inciertas.[8] Cuando los hechos les niegan la comprensión y la confianza que anhelan, recurren a su intuición para que esta se las proporcione. La negación de la ignorancia es tanto más tentadora cuanto mayor es la ignorancia.

La negación de la ignorancia también explica otro rompecabezas. Ante pruebas como las que aquí hemos presentado, muchos líderes sacan una conclusión aparentemente paradójica. Sus decisiones basadas en el instinto pueden no ser perfectas, argumentan, pero, si las alternativas más sistemáticas también están lejos de ser perfectas, no vale la pena adoptarlas. Recordemos, por ejemplo, que la correlación media entre las calificaciones de los juzgadores humanos y el rendimiento de los empleados es de .28 (PC = 59 %). Según el mismo estudio, y en consonancia con las pruebas que hemos revisado, la predicción mecánica podría ser mejor, pero no en mucho: su exactitud predictiva es de .44 (PC = 65 %). Un ejecutivo podría preguntarse: ¿por qué molestarse?

La respuesta es que, en algo tan importante como las decisiones de a quién contratar, este incremento de la validez tiene un enorme valor. Los mismos ejecutivos realizan habitualmente cambios significativos en su forma de trabajar para obtener ganancias que no son tan grandes. Racionalmente entienden que el éxito nunca puede estar garantizado y que lo que tratan de lograr con sus decisiones es una mayor probabilidad de éxito. También entienden la probabilidad. Ninguno de ellos compraría un número de lotería con un 59 por ciento de posibilidades de ganar si pudiera comprar, por el mismo precio, uno con un 65 por ciento de posibilidades.

El problema es que el «precio» en esta situación no es el mismo. El juicio intuitivo tiene su recompensa, que es la señal interna. La gente está

preparada para confiar en un algoritmo que alcance un nivel de exactitud muy alto porque le da una sensación de certeza que iguala o excede la que le proporciona la señal interna.[9] Sin embargo, renunciar a la recompensa emocional de la señal interna es un precio muy alto cuando la alternativa es algún tipo de proceso mecánico que ni siquiera pretende tener una gran validez.

Esta observación tiene una importante implicación para la mejora del juicio. A pesar de todas las pruebas a favor de los métodos de predicción mecánicos y algorítmicos, y a pesar del cálculo racional que muestra con claridad el valor de las mejoras incrementales en la exactitud de la predicción, muchos responsables de la toma de decisiones rechazarán los métodos de la toma de decisiones que les privan de la capacidad de ejercer su intuición. Mientras los algoritmos no sean casi perfectos —y, en muchos ámbitos, la ignorancia objetiva dicta que nunca lo serán—, el juicio humano no será sustituido. Por eso hay que mejorarlo.

SOBRE LA IGNORANCIA OBJETIVA

- «Donde hay predicción, hay ignorancia, y probablemente más de la que pensamos. ¿Hemos comprobado si los expertos en los que confiamos aciertan más que los chimpancés que lanzan dardos?»
- «Cuando confías en tu instinto por una señal interna, y no por algo que realmente sabes, estás negando tu ignorancia objetiva.»
- «Los modelos lo hacen mejor que las personas, pero no mucho más. Comúnmente encontramos juicios humanos mediocres y modelos un poco mejores. Aun así, lo mejor es bueno, y los modelos son mejores.»
- «Puede que nunca nos sintamos cómodos utilizando un modelo para tomar estas decisiones, pues necesitamos la señal interna para tener suficiente confianza. Por lo tanto, hemos de asegurarnos de tener el mejor proceso de decisión posible.»

12
El valle de lo normal

Pasemos ahora a una cuestión más amplia: ¿cómo lograr comodidad en un mundo en el que muchos problemas son fáciles de resolver, pero muchos otros están dominados por la ignorancia objetiva? Cuando la ignorancia objetiva es seria, pasado un tiempo deberíamos tomar conciencia de la inutilidad de las bolas de cristal en los asuntos humanos. Sin embargo, esa no es nuestra experiencia habitual del mundo. En cambio, como se ha indicado en el capítulo anterior, mantenemos una voluntad sin reservas de hacer audaces predicciones sobre el futuro partiendo de muy escasa información útil. En este capítulo, abordaremos la sensación predominante y errónea de que los acontecimientos que no han podido predecirse pueden, no obstante, comprenderse.

¿Qué significa realmente esta creencia? Planteamos esta cuestión en dos contextos: el de las ciencias sociales y el de la experiencia del acontecer de la vida cotidiana.

PREDICCIÓN DE TRAYECTORIAS VITALES

En 2020, un nutrido grupo de ciento doce investigadores dirigidos por Sara McLanahan y Matthew Salganik, profesores de Sociología de la Universidad de Princeton, publicaron un artículo excepcional en *Proceedings of the National Academy of Sciences*.[1] Los investigadores trataban de averiguar hasta qué punto los científicos sociales entienden en verdad lo que ocurrirá en las trayectorias vitales de las familias socialmente vulnerables. Sabiendo lo que saben, ¿hasta qué punto pueden los científicos sociales predecir los acontecimientos de la vida de una familia? En concreto, ¿qué nivel de exactitud pueden alcanzar los expertos a la hora de predecir acontecimientos de la vida de estas familias utilizando la información que los sociólogos suelen reunir y aplicar en sus investigaciones? Empleando nuestros términos, el objetivo del estudio era

medir el nivel de ignorancia objetiva que permanece en estos acontecimientos después de que los sociólogos hayan hecho su trabajo.

Los autores utilizaron material del «Fragile Families and Child Wellbeing Study», una investigación longitudinal a gran escala en niños a los que se hizo un seguimiento desde el nacimiento hasta los quince años de edad. La enorme base de datos contenía varios miles de ítems de información sobre las familias de casi cinco mil niños, la mayoría de ellos nacidos de padres no casados en grandes ciudades de Estados Unidos. Los datos abarcaban temas como la educación y el empleo de los abuelos de los niños, detalles sobre la salud de todos los miembros de la familia, índices de situación económica y social, respuestas a múltiples cuestionarios y test de aptitud cognitiva y de personalidad. Una información extraordinariamente rica de la que los científicos sociales han hecho un buen uso: se han escrito más de setecientos cincuenta artículos científicos basados en los datos del estudio de familias vulnerables. Muchos de estos trabajos utilizaron los datos sobre los niños y sus familias para explicar los resultados de su vida, como calificaciones de la escuela secundaria y antecedentes penales.

El estudio dirigido por el equipo de Princeton se centró en la predictibilidad de seis resultados observados cuando el niño tenía quince años, entre ellos un desahucio reciente, el promedio de notas del niño y una medición general de las circunstancias materiales de su hogar. Los organizadores utilizaron lo que llamaron «método de la tarea común». Invitaron a los equipos de investigadores a competir en la creación de predicciones exactas de los seis resultados elegidos utilizando la masa de datos disponibles sobre cada familia en el estudio de las familias vulnerables. Esta clase de reto es nuevo en las ciencias sociales, pero habitual en las ciencias informáticas, donde se suele invitar a los equipos a competir en tareas como la traducción automática de un conjunto estándar de textos o la detección de un animal en un gran conjunto de fotografías. Los logros del equipo ganador de estas competiciones definen la tecnología de última generación en un momento dado, que siempre es superada en la siguiente competición. En una tarea de predicción propia de las ciencias sociales, en la que no se espera una mejora rápida, es razonable utilizar la predicción más exacta obtenida en la competición como medida de predictibilidad del resultado a partir de estos datos (en otras palabras: del nivel residual de ignorancia objetiva).

El reto despertó un considerable interés entre los investigadores. El informe final presentó los resultados de ciento sesenta equipos altamente cualificados procedentes de un elenco internacional de candidatos mucho más grande. La mayoría de los competidores seleccionados se presentaron como científicos que operaban con datos y utilizaban el aprendizaje automático.

En la primera fase de la competición, los equipos participantes tuvieron acceso a todos los datos de la mitad de la muestra total; los datos incluían los seis resultados.[2] Utilizaron estos «datos de entrenamiento» para entrenar un algoritmo predictivo. A continuación, sus algoritmos se aplicaron a una muestra de familias que no se había utilizado para entrenar el algoritmo. Los investigadores midieron la exactitud usando el ECM: el error de predicción de cada caso era el cuadrado de la diferencia entre el resultado real y la predicción del algoritmo.

¿Cuáles fueron los modelos ganadores? Los complejos algoritmos de aprendizaje automático entrenados en un gran conjunto de datos superaron, por supuesto, las predicciones de los modelos lineales simples (y, por extensión, a los juzgadores humanos). Sin embargo, la mejora de los modelos de IA con respecto a un modelo muy simple fue ligera y su exactitud predictiva siguió siendo decepcionantemente baja. En la predicción de desahucios, el mejor modelo alcanzó una correlación de .22 (PC = 57 %).[3] Se encontraron resultados similares para otros acontecimientos, como si el cuidador principal había sido despedido o había recibido formación para su trabajo y qué puntuación obtendría el niño en una medición, evaluada por él mismo, de «agallas» (un rasgo de personalidad que combina la perseverancia y la pasión por una meta concreta). Para estos casos, las correlaciones se situaron entre .17 y .24 (PC = 55-58 %).

Dos de los seis resultados finales eran agregados, mucho más predecibles. Las correlaciones predictivas fueron de .44 (PC = 65 %) con el promedio de notas del niño y de .48 (PC = 66 %) con una medición resumida de las dificultades materiales durante los doce meses anteriores. Esta medición se basó en once preguntas, entre las que se incluían algunas como: «¿Ha pasado hambre alguna vez?» o «¿Le cortaron el teléfono?». Es bien sabido que las mediciones sumadas son a la vez más predictivas y más predecibles que las mediciones de resultados simples. La principal conclusión del reto

es que una gran masa de información predictiva no es suficiente para la predicción de acontecimientos concretos en la vida de las personas (y hasta la predicción de agregados es bastante limitada).

Los resultados observados en esta investigación son bastante típicos y muchas de las correlaciones que los científicos sociales presentan están en este rango. Un amplio examen de la investigación en psicología social, que abarcó veinticinco mil estudios con ocho millones de sujetos durante cien años, concluyó que «los efectos psicológicos sociales arrojan por regla general un valor de *r* (coeficiente de correlación) igual a .21».[4] Correlaciones mucho más altas, como la antes mencionada de .60 entre la estatura de los adultos y el tamaño de los pies, son comunes cuando se trata de mediciones físicas, pero muy raras en las ciencias sociales. Una revisión de setecientos ocho estudios de las ciencias conductuales y cognitivas encontró que solo el 3 por ciento de las correlaciones registradas eran de .50 o más.[5]

Estos coeficientes de correlación tan bajos pueden resultar sorprendentes a quien esté acostumbrado a leer sobre resultados que se presentan como «estadísticamente significativos» o incluso «altamente significativos». Los términos estadísticos suelen ser engañosos para el lector no especializado, y «significativos» puede ser el peor ejemplo de ello. Cuando un hallazgo se califica de «significativo», no debemos concluir que el efecto que describe es convincente. Solo significa que es poco probable que el descubrimiento sea producto únicamente del azar. Con una muestra lo bastante grande, una correlación puede ser a la vez muy «significativa» y demasiado pequeña para que merezca ser analizada.

La limitada predictibilidad de los resultados simples en el estudio del reto acarrea un preocupante mensaje sobre la diferencia entre comprensión y predicción. El estudio sobre las familias vulnerables se considera un tesoro de la ciencia social y, como hemos visto, sus datos han sido utilizados en un vasto cuerpo de investigación. Los estudiosos que llevaron a cabo ese trabajo seguramente sintieron que avanzaban en la comprensión de la vida de las familias vulnerables. Por desgracia, a esta sensación de progreso no le acompañó la capacidad de hacer predicciones granulares sobre acontecimientos individuales en las vidas individuales. El resumen introductorio del informe de varios autores sobre el reto de las familias vulnerables contenía una dura advertencia: «Los investigadores deben

reconciliar la idea de que entienden las trayectorias de la vida con el hecho de que ninguna de las predicciones fue muy acertada».[6]

COMPRENSIÓN Y PREDICCIÓN

La lógica que subyace a esta conclusión pesimista requiere cierta elaboración. Cuando los autores del reto de las familias vulnerables equiparan la comprensión con la predicción (o la ausencia de una con la ausencia de la otra), utilizan el término «comprensión» en un sentido específico. Hay otras acepciones de la palabra: quien dice que comprende un concepto matemático, o que comprende lo que es el amor, no está aludiendo a alguna capacidad para hacer predicciones concretas.

Sin embargo, en el discurso de las ciencias sociales, y en la mayoría de las conversaciones cotidianas, la pretensión de comprender algo lo es de comprender su «causa». Los sociólogos que recogieron y estudiaron los miles de variables del «Fragile Families and Child Wellbeing Study» buscaban las causas de los resultados que observaban. Los médicos que comprenden lo que aqueja a un paciente afirman que la patología que han diagnosticado es la causa de los síntomas que han observado. «Comprender» es describir una cadena causal.[7] La capacidad de hacer una predicción es una medición de si dicha cadena causal ha sido de hecho identificada. Y la correlación —la medición de la exactitud de la predicción — es una medición de la causalidad que podemos explicar.

Esta última afirmación puede sorprender a quien haya tratado con la estadística elemental y recuerde la advertencia, tantas veces repetida, de que «la correlación no implica causalidad». Consideremos, por ejemplo, la correlación entre la talla de los zapatos y las aptitudes matemáticas de los niños: obviamente, una variable no es la causa de la otra. La correlación surge del hecho de que tanto la talla de los zapatos como los conocimientos matemáticos aumentan con la edad del niño. La correlación es real y apoya una predicción: si se sabe que un niño tiene los pies grandes, habría que predecir un nivel de matemáticas más alto que si se sabe que el niño tiene los pies pequeños. Pero no hay que deducir una relación causal de esta correlación.

Sin embargo, debemos recordar que, si bien la correlación no implica causalidad, esta «sí» implica correlación. Cuando hay una relación causal,

deberíamos encontrar una correlación. Si no se encuentra ninguna correlación entre la edad y la talla de los zapatos en los adultos, se puede concluir con seguridad que, tras el final de la adolescencia, la edad no hace que los pies aumenten de tamaño y que hay que buscar en otra parte las causas de las diferencias en la talla de los zapatos.

Donde hay causalidad, hay correlación. De ello se desprende que, cuando hay causalidad, deberíamos ser capaces de predecir, y la correlación —la exactitud de esta predicción— es una medición del grado de causalidad que entendemos. De ahí que la conclusión de los investigadores de Princeton sea la siguiente: el grado en que los sociólogos pueden predecir sucesos como los desahucios, medido por una correlación de .22, es una indicación de lo mucho —o lo poco— que entienden sobre las trayectorias vitales de esas familias. La ignorancia objetiva no solo pone un techo a nuestras predicciones, sino también a nuestra comprensión.

Entonces ¿qué quieren decir la mayoría de los profesionales cuando afirman con seguridad que entienden de su campo? ¿Cómo pueden pronunciarse sobre las causas de los fenómenos que observan y ofrecer predicciones seguras sobre ellos? En resumen, ¿por qué los profesionales —como todos nosotros— parecen subestimar nuestra ignorancia objetiva del mundo?

PENSAMIENTO CAUSAL

Si, al leer la primera parte de este capítulo, el lector se preguntó qué es lo que conduce a los desahucios y a otros resultados de la vida en las familias vulnerables, no hizo sino participar del mismo tipo de pensamiento que el de los investigadores cuyos intentos hemos descrito. Aplicó el «pensamiento estadístico»: se interesó por los conjuntos, como la población de familias vulnerables, y por las estadísticas que los describen, incluidos los promedios, las varianzas, las correlaciones, etcétera. No se centró en casos individuales.

Otro modo de pensar que brota de forma más natural en nuestras mentes se denominará aquí «pensamiento causal».[8] Este último crea relatos en los que sucesos, personas y objetos específicos se afectan unos a otros. Para experimentar el pensamiento causal, imagínese a un trabajador social que sigue el caso de muchas familias desfavorecidas. Acaba de enterarse de que

una de estas familias, los Jones, ha sido desahuciada. Su reacción a este hecho se basa en lo que sabe de los Jones. Resulta que Jessica Jones, el sostén de la familia, fue despedida hace unos meses. No pudo encontrar otro trabajo y, desde entonces, no ha podido pagar todo el alquiler. Hizo pagos parciales, varias veces suplicó al administrador del edificio e incluso pidió al trabajador social que interviniera (este lo hizo, pero el administrador permaneció impasible). En este contexto, el desahucio de los Jones es triste, pero no sorprendente. Se siente, de hecho, como el final lógico de una cadena de acontecimientos, el desenlace inevitable de una tragedia que estaba escrita.

Cuando nos rendimos ante esta sensación de inevitabilidad, perdemos de vista lo fácil que podría haber sido todo de otra manera, cómo, en cada bifurcación del camino, el destino podría haber tomado un derrotero diferente. Jessica podría haber conservado su trabajo. Podría haber encontrado rápidamente otro. Un familiar podría haber acudido en su ayuda. El trabajador social podría haber sido un defensor más eficaz. El administrador podría haber sido más comprensivo y haber concedido a la familia unas semanas de respiro, lo que hubiera hecho posible que Jessica encontrara un trabajo y se pusiera al día con el alquiler.

Estos relatos alternativos son tan poco sorprendentes como el primero (si se conoce el final). Sea cual fuere el resultado (desahucio o no), una vez que se ha producido, el pensamiento causal hace que parezca totalmente explicable, incluso predecible.

COMPRENDIENDO EL VALLE DE LO NORMAL

Esta observación tiene una explicación psicológica. Algunos acontecimientos son sorprendentes: una pandemia mortal, un ataque contra las Torres Gemelas, un fondo de inversión estrella que resulta ser un esquema Ponzi. También en nuestra vida personal hay sucesos imprevistos: enamorarse de una persona desconocida, la muerte repentina de un hermano pequeño, una herencia inesperada. Otros acontecimientos se esperan seriamente, como el regreso de la escuela a la hora establecida de un alumno de segundo de primaria.

Sin embargo, la mayor parte de la experiencia humana se sitúa entre estos dos extremos. A veces nos encontramos en un estado en el que esperamos

un acontecimiento concreto, y a veces nos encontramos con sorpresas. Pero la mayoría de las cosas ocurren en el amplio valle de lo normal, donde los acontecimientos no son del todo esperados ni particularmente sorprendentes. En este momento, por ejemplo, el lector no tiene ninguna expectativa concreta sobre lo que viene en el siguiente párrafo. Se sorprendería si de repente escribiéramos en turco, pero hay una gran variedad de cosas que podríamos decir sin causarle gran sorpresa.

En el valle de lo normal, los acontecimientos se desarrollan como en el desahucio de los Jones: parecen normales después de producirse, aunque no se esperaban y aunque no podríamos haberlos previsto. Esto se debe a que el proceso de comprensión de la realidad es retrospectivo.[9] Un suceso que no fue anticipado seriamente (el desahucio de la familia Jones) desencadena la búsqueda en la memoria de una causa candidata (el difícil mercado de trabajo, el administrador inflexible). La búsqueda se detiene cuando se da con un buen relato. Si el resultado fuera el contrario, la búsqueda habría dado con causas igualmente convincentes (la tenacidad de Jessica Jones, el administrador comprensivo).

Como ilustran estos ejemplos, muchos acontecimientos de una historia normal se explican literalmente por sí mismos. Se habrá notado que, en las dos versiones de la historia del desahucio, el administrador del edificio no era la misma persona: el primero era indolente y el segundo era compasivo. Sin embargo, la única pista sobre el carácter del administrador fue su comportamiento, que demostraba justamente su carácter. Teniendo en cuenta lo que ahora sabemos de él, su comportamiento parece coherente. Es el propio acontecimiento el que nos revela su causa.

Cuando se explica así un resultado inesperado, pero no sorprendente, el desenlace siempre tiene sentido. Esto es lo que entendemos por «comprender» una historia, y esto es lo que hace que la realidad parezca predecible (en retrospectiva). Dado que el acontecimiento se explica por sí mismo a medida que se produce, se tiene la impresión de que se podría haber anticipado.

En términos más generales, nuestra comprensión del mundo depende de nuestra extraordinaria capacidad para construir relatos que expliquen los acontecimientos que observamos. La búsqueda de causas casi siempre tiene éxito, porque estas pueden extraerse de una reserva ilimitada de hechos y creencias sobre el mundo. Como sabe cualquiera que escuche las noticias

de la noche, por ejemplo, ningún gran movimiento del mercado de valores queda jamás sin explicación. El mismo flujo de noticias puede «explicar» tanto una caída de los índices (¡los inversores, nerviosos, están preocupados por las noticias!) como una subida (¡los inversores siguen mostrándose optimistas!).

Cuando la búsqueda de una causa obvia fracasa, nuestro primer recurso es dar una explicación llenando un espacio en blanco en nuestro modelo del mundo. Introducimos entonces un hecho que no conocíamos antes (el administrador era una persona inusualmente compasiva). Solo cuando nuestro modelo del mundo no puede modificarse para producir el resultado, etiquetamos este como sorprendente y empezamos a buscar una explicación más elaborada de él. La verdadera sorpresa solo se produce cuando la retrospectiva rutinaria falla.

Esta continua interpretación causal de la realidad es nuestra forma de «comprender» el mundo. Nuestra sensación de comprender la vida a medida que se desarrolla consiste en un flujo continuo de retrospección en el valle de lo normal. Esta sensación es fundamentalmente causal: los nuevos acontecimientos, una vez conocidos, eliminan las alternativas, y el relato deja poco espacio a la incertidumbre. Como sabemos por la investigación clásica sobre la retrospección, aunque la incertidumbre subjetiva existe durante un tiempo, los recuerdos de esta se borran en gran medida cuando la incertidumbre se resuelve.[10]

DENTRO Y FUERA

Hemos deslindado dos modos de pensar los acontecimientos: el estadístico y el causal. Al permitirnos categorizar los acontecimientos en tiempo real como normales o anormales, el modo causal nos ahorra mucho esfuerzo. Los acontecimientos anormales demandan al instante un costoso esfuerzo de búsqueda de información relevante, tanto en el entorno como en la memoria. La expectativa activa —esperar atentamente a que ocurra algo— también requiere un esfuerzo. En cambio, el flujo de acontecimientos en el valle de lo normal requiere un escaso trabajo mental. Es posible que nuestro vecino sonría cuando se cruce con nosotros, o que parezca preocupado y se limite a saludar con la cabeza; ninguno de estos hechos nos llamará mucho la atención si ambos han sido razonablemente frecuentes en el pasado. Si la

sonrisa es excepcionalmente amplia o el saludo con la cabeza excepcionalmente leve, es posible que busquemos en nuestra memoria una causa posible. El pensamiento causal evita esfuerzos innecesarios, al tiempo que mantiene la vigilancia necesaria para detectar aconteceres anormales.

El pensamiento estadístico, en cambio, exige un esfuerzo. Requiere los recursos de atención que el Sistema 2, asociado al pensamiento lento y deliberado, puede aportar.[11] Más allá de un nivel elemental, el pensamiento estadístico también exige una formación especializada. Este tipo de pensamiento comienza con conjuntos y considera los casos individuales como elementos de categorías más amplias. El desahucio de los Jones no se ve como el resultado de una cadena de acontecimientos específicos, sino como un resultado estadísticamente probable (o improbable), dadas las observaciones previas de casos que comparten características predictivas con el de los Jones.

La distinción entre estos dos puntos de vista es un tema recurrente en este libro. Confiar en el pensamiento causal para un solo caso es una fuente de errores predecibles. Adoptar el punto de vista estadístico, que también llamaremos la «visión externa», es una forma de evitar estos errores.

Llegados a este punto, lo único que necesitamos subrayar es que el modo causal nos resulta mucho más natural. Incluso las explicaciones que deberían tratarse propiamente como estadísticas se convierten fácilmente en relatos causales. Considérense afirmaciones como «Fracasaron porque les faltaba experiencia» o «Tuvieron éxito porque tenían un líder brillante». Sería fácil pensar en contraejemplos: casos en los que equipos inexpertos tuvieron éxito y líderes brillantes fracasaron. Las correlaciones de experiencia y brillantez con el éxito son, en el mejor de los casos, moderadas y probablemente bajas. Sin embargo, la atribución causal no resulta difícil de hacer. Cuando la causalidad es plausible, nuestra mente convierte con facilidad una correlación, por baja que sea, en una fuerza causal y explicativa. Se acepta el liderazgo brillante como explicación satisfactoria del éxito y la inexperiencia como explicación del fracaso.

La confianza en estas explicaciones injustificadas es seguramente inevitable si la alternativa es dejar de entender nuestro mundo. Sin embargo, el pensamiento causal y la ilusión de comprender el pasado contribuyen a las predicciones del futuro excesivamente confiadas. Como veremos, la preferencia por el pensamiento causal contribuye también a

descuidar el ruido como fuente de error, porque la noción de ruido es sobre todo estadística.

El pensamiento causal nos ayuda a dar sentido a un mundo que es mucho menos predecible de lo que creemos. También explica por qué vemos el mundo como mucho más predecible de lo que realmente es. En el valle de lo normal no hay sorpresas ni inconsistencias. El futuro parece tan predecible como el pasado. Y el ruido ni se oye ni se ve.

SOBRE LOS LÍMITES DE LA COMPRENSIÓN

- «Correlaciones de alrededor de .20 (PC = 56%) son bastante comunes en los asuntos humanos.» ■ «La correlación no implica causalidad, pero esta sí implica correlación.»
- «La mayoría de los acontecimientos normales ni se esperan ni sorprenden, y no necesitan ninguna explicación.»
- «En el valle de lo normal, los acontecimientos ni se esperan ni sorprenden; se explican por sí mismos.»
- «Creemos que entendemos lo que está pasando, pero ¿podríamos haberlo predicho?»

CUARTA PARTE

Cómo se genera el ruido

¿Cuál es el origen del ruido? ¿Y del sesgo? ¿Qué mecanismos mentales originan la variabilidad de nuestros juicios y los errores compartidos que los afectan? En resumen, ¿qué sabemos sobre la psicología del ruido? Estas son las cuestiones de las que trataremos ahora.

En primer lugar, describiremos cómo algunas de las operaciones del pensamiento rápido del «Sistema 1» son responsables de muchos errores de juicio. En el capítulo 13, presentaremos tres importantes heurísticas del juicio en las que se basa ampliamente el Sistema 1. Demostraremos cómo estas heurísticas causan errores predecibles, direccionales (sesgo estadístico), así como ruido.

El capítulo 14 se centrará en el emparejamiento —una operación particular del Sistema 1— y analizará los errores que puede producir.

En el capítulo 15, nos ocuparemos de un complemento indispensable en todos los juicios: la escala sobre la que se hacen estos últimos. Demostraremos que la elección de una escala adecuada es un requisito previo para el buen juicio y que las escalas mal definidas o inadecuadas son una fuente importante de ruido.

En el capítulo 16 exploraremos el origen psicológico del que puede ser el tipo de ruido más interesante: los patrones de respuesta que tienen personas diferentes a casos distintos. Como las personalidades individuales, estos patrones no son aleatorios, y en su mayoría son estables en el tiempo, pero sus efectos no son muy predecibles.

Finalmente, en el capítulo 17, resumiremos lo que hemos aprendido sobre el ruido y sus componentes. Esta investigación nos llevará a proponer una respuesta al enigma que antes planteamos: ¿por qué el ruido, a pesar de su omnipresencia, rara vez se considera un problema importante?

13
Heurísticas, sesgos y ruido

Este libro amplía medio siglo de investigación sobre el juicio humano intuitivo, el llamado «programa de heurísticas y sesgos». Las primeras cuatro décadas de este programa de investigación se revisaron en *Pensar rápido, pensar despacio*, que examinaba los mecanismos psicológicos que explican tanto las maravillas como los defectos del pensamiento intuitivo. [1] La idea central del programa era que las personas a las que se les hace una pregunta difícil utilizan operaciones de simplificación, llamadas «heurísticas». En general, las heurísticas, producidas por el pensamiento rápido e intuitivo, también conocido como «pensamiento del Sistema 1», son bastante útiles y ofrecen respuestas adecuadas. Sin embargo, a veces dan lugar a sesgos, que hemos descrito como «errores de juicio sistemáticos y predecibles.

El programa de heurísticas y sesgos se centraba en lo que las personas tienen en común, no en lo que difieren. Se demostró que los procesos que provocan los errores de juicio son ampliamente compartidos. En parte debido a esta historia, las personas que están familiarizadas con la noción de «sesgo psicológico» a menudo suponen que este siempre produce un «sesgo estadístico», un término que utilizamos en este libro para referirnos a las mediciones o juicios que generalmente se desvían de la verdad en la misma dirección. De hecho, los sesgos psicológicos crean un sesgo estadístico cuando son ampliamente compartidos. Sin embargo, los sesgos psicológicos crean ruido en el sistema cuando los juzgadores están mentalmente sesgados de distintas maneras o en diferente medida. Con independencia de que provoquen sesgo estadístico o ruido, por supuesto, los sesgos psicológicos siempre generan errores.

DIAGNÓSTICO DE LOS SESGOS

Los sesgos de los juicios suelen identificarse en relación con un valor real. Existe un sesgo en los juicios predictivos si los errores van en su mayoría

en una dirección y no en otra. Por ejemplo, cuando las personas prevén el tiempo que les llevará completar un proyecto, la media de sus estimaciones suele ser muy inferior al tiempo que realmente necesitarán. A este conocido sesgo psicológico se le suele llamar «falacia de la planificación».

Sin embargo, a menudo no existe un valor real con el que poder comparar los juicios. Teniendo en cuenta lo mucho que hemos insistido en que el sesgo estadístico solo puede detectarse cuando se conoce el valor real, quizá alguien se pregunte cómo pueden estudiarse los sesgos psicológicos cuando se desconoce la verdad. La respuesta es que los investigadores confirman un sesgo psicológico al observar que un factor que no debería afectar al juicio tiene un efecto estadístico sobre él, o que un factor que debería afectar al juicio no lo tiene.

Para ilustrar este método, podemos volver al ejemplo del campo de tiro. Imaginemos que los equipos A y B han efectuado sus disparos y que estamos mirando la parte posterior del blanco (figura 12). En este ejemplo, no se sabe dónde está la diana (se desconoce el valor real). Por tanto, se desconoce cuál es el sesgo de los dos equipos en relación con el centro del blanco. Sin embargo, podemos detectar que, en el panel 1, los dos equipos estaban apuntando a la misma diana y que, en el panel 2, el equipo A estaba apuntando a una diana y el equipo B a otra diferente.

FIGURA 12. Una vista del dorso del blanco en un experimento para detectar el sesgo

A pesar de la ausencia de un blanco, ambos paneles ofrecen pruebas de un sesgo sistemático. En el panel 1, los disparos de los dos equipos difieren,

aunque deberían ser idénticos. Este patrón se asemeja a lo que se vería en un experimento en el que dos grupos de inversores leyeran planes de negocio sustancialmente idénticos, pero impresos en un tipo de letra diferente y en un papel distinto. Si estos detalles irrelevantes marcan una diferencia en el juicio de los inversores, existe un sesgo psicológico. No sabemos si los inversores que quedaron impresionados por el tipo de letra elegante y el papel satinado son demasiado positivos o si los que leyeron la versión más tosca son demasiado negativos. Sin embargo, sabemos que sus juicios son diferentes, aunque no deberían serlo.

El panel 2 ilustra el fenómeno contrario. Como los equipos apuntaban a blancos diferentes, los impactos deberían ser distintos, pero están centrados en el mismo punto. Imagínese, por ejemplo, que a dos grupos de personas se les hace la misma pregunta que en el capítulo 4 se le hizo al lector sobre Michael Gambardi, pero con un giro. A un grupo se le pide, como al lector, que estime la probabilidad de que Gambardi siga en su puesto de trabajo dentro de dos años; al otro se le pide que calcule la probabilidad de que siga en su puesto de trabajo dentro de tres años. Los dos grupos deberían llegar a conclusiones diferentes, porque es evidente que hay más formas de perder el empleo en tres años que en dos. Sin embargo, las pruebas sugieren que las estimaciones de probabilidad de los dos grupos diferirán poco, si es que lo hacen.[2] Las respuestas deberían ser muy diferentes, pero no lo son, lo cual indica que se ignora un factor que tendría que influir en los juicios. (Este sesgo psicológico se llama «insensibilidad al alcance».)

Los errores de juicio sistemáticos se han demostrado en muchos campos, y el término «sesgo» se utiliza ahora en numerosos ámbitos, como el de los negocios, la política, las normas públicas y privadas y el derecho. Tal y como se emplea comúnmente la palabra, su significado es amplio. Además de la definición cognitiva que aquí utilizamos (referida a un mecanismo psicológico y al error que este suele producir), la palabra se usa con frecuencia para indicar que alguien tiene prejuicios contra un determinado grupo (por ejemplo, los relativos al sexo o la raza). También puede significar que alguien está a favor de una conclusión concreta, como cuando leemos que alguien está mentalmente sesgado por un conflicto de intereses o por una opinión política. Incluimos estos tipos de sesgo en nuestra discusión sobre la psicología de los errores de juicio porque todos los sesgos psicológicos provocan tanto sesgo estadístico como ruido.

Hay un uso al que nos oponemos con firmeza. En este uso, se atribuyen costosos fracasos a un «sesgo» no especificado, y el reconocimiento del error va acompañado de la promesa de «trabajar duro para eliminar los sesgos en nuestra toma de decisiones». Estas declaraciones no significan más que «se cometieron errores» y «nos esforzaremos por mejorar». No cabe duda de que algunos fallos se deben en realidad a errores predecibles asociados a sesgos psicológicos específicos, y creemos en la viabilidad de las intervenciones para reducir los sesgos (y el ruido) en juicios y decisiones. Sin embargo, culpar de todos los resultados indeseados a los sesgos es una explicación inútil. Recomendamos reservar la palabra «sesgo» para errores específicos e identificables y los mecanismos que los producen.

SUSTITUCIÓN

Para experimentar el proceso heurístico, el lector puede responder a la siguiente pregunta, que ilustra varios temas esenciales acerca de las heurísticas y los sesgos. Como de costumbre, sacará más provecho del ejemplo si ofrece sus propias respuestas.

Bill tiene treinta y tres años. Es inteligente, pero poco imaginativo, compulsivo y, en general, apático. En la escuela era bueno en matemáticas, pero flojo en ciencias sociales y lenguas.

Lo que sigue es una lista de ocho posibilidades referentes a la vida actual de Bill.

Lea la lista y seleccione las dos que considere más probables.

- Bill es un médico que juega al póquer por afición.
- Bill es arquitecto.
- Bill es contable.
- Bill toca jazz por afición.
- Bill practica el surf por afición.
- Bill es periodista.
- Bill es un contable que toca jazz por afición.
- Bill escala montañas por afición.

Ahora, vuelva a leer la lista y seleccione las dos categorías en las que Bill más se parece a una persona típica de esa categoría. Puede elegir las mismas o diferentes que antes.

Estamos casi seguros de que ha elegido las mismas categorías como las más altas en probabilidad y en parecido. La razón de nuestra seguridad es que múltiples experimentos han demostrado que la gente da respuestas idénticas a las dos cuestiones.[3]

Sin embargo, la similitud y la probabilidad son, en realidad, muy diferentes. Por ejemplo, pregúntese cuál de las siguientes afirmaciones tiene más sentido.

- Bill se ajusta a la idea que tengo de una persona que toca jazz por afición.
- Bill se ajusta a la idea que tengo de un contable que toca jazz por afición.

Ninguna de estas afirmaciones es buena, pero una de ellas es claramente menos deslucida que la otra. Bill tiene más en común con un contable que toca jazz por afición que con una persona que toca jazz por afición. Ahora considere lo siguiente: ¿cuál de estas dos opciones es más probable?

- Bill toca jazz por afición.
- Bill es un contable que toca jazz por afición.

Podrá estar tentado de elegir la segunda respuesta, pero la lógica no se lo permite. La probabilidad de que Bill toque jazz por afición «tiene que ser» mayor que la probabilidad de que sea un contable que toque jazz. Recuerde los diagramas de Venn. Si Bill toca jazz y es contable, sin duda toca jazz. Añadir detalles a una descripción solo puede hacerla menos probable, aunque puede hacerla más representativa, como en el presente caso.

La teoría de la heurística del juicio propone que las personas usan a veces la respuesta a una pregunta más fácil para responder a la más difícil. Entonces ¿qué pregunta es más fácil de responder?: «¿En qué medida se parece Bill a un típico jazzista aficionado?» o «¿Qué probabilidad hay de que Bill sea un músico de jazz aficionado?». La pregunta sobre la similitud es, sin discusión, más fácil, lo que hace probable que sea la que la gente responda cuando se le pide que estime la probabilidad.

El lector habrá ya conocido la idea esencial del programa de heurísticas y sesgos: una heurística para responder a una pregunta difícil es encontrar la respuesta a una pregunta más fácil. La sustitución de una pregunta por la otra causa errores predecibles llamados «sesgos psicológicos».

Este tipo de sesgo es patente en el ejemplo de Bill. Cuando se sustituye un juicio de similitud por un juicio de probabilidad, es inevitable que se produzcan errores, ya que la probabilidad se halla limitada por una lógica especial.

En particular, los diagramas de Venn solo se aplican a la probabilidad, no a la similitud. De ahí el previsible error lógico que muchas personas cometen.

Para otro ejemplo de una propiedad estadística ignorada, el lector puede recordar cómo pensó en la pregunta sobre Gambardi del capítulo 4. Si es como la de la mayoría de las personas, su evaluación de las posibilidades de éxito de Michael Gambardi se basó enteramente en lo que el caso le decía sobre él. Luego intentó hacer coincidir su descripción con la imagen de un director ejecutivo (CEO) con éxito.

¿Se le ocurrió considerar la probabilidad de que un CEO elegido al azar siga ocupando el mismo puesto dos años después? Probablemente no.

Habrá entendido esta «información de tasa-base» como una medida de la dificultad de permanecer como CEO. Si este enfoque le parece extraño, piense en cómo estimaría la probabilidad de que un alumno concreto apruebe un examen. Sin duda, la proporción de estudiantes que suspenden el examen es relevante, ya que da una indicación de lo difícil que este es. Del mismo modo, la tasa-base de permanencia de los CEO es relevante para el problema de Gambardi.

Ambas preguntas son ejemplos de lo que hemos llamado «visión externa»:cuando se adopta este punto de vista, se piensa en el alumno, o en Gambardi, como miembro de una clase de casos similares. Se piensa estadísticamente en la clase en lugar de pensar causalmente en el caso principal.

La visión externa puede suponer una gran diferencia y evitar errores importantes. Unos minutos de investigación revelarían que las estimaciones sobre el reemplazo de los CEO en las empresas estadounidenses rondan el 15 por ciento anual.[4] Esta estadística indica que el CEO medio entrante tiene una probabilidad aproximada del 72 por ciento de seguir en el cargo

dos años después. Por supuesto, esta cifra solo es un punto de partida, y los detalles del caso de Gambardi afectarán a la estimación final que el lector hará sobre él. Sin embargo, si se centró solo en lo que se le dijo sobre Gambardi, ignoró una pieza clave de información. (Una revelación: inventamos el caso Gambardi para ilustrar un juicio ruidoso; tardamos semanas en darnos cuenta de que también era un buen ejemplo del sesgo que describimos aquí, y que se denomina «ignorancia de la tasa-base». Pensar en las tasas-base no es más automático para los autores de este libro que para cualquier otra persona.)

La sustitución de una cuestión por otra no se limita a la similitud y la probabilidad. Otro ejemplo es la sustitución de un juicio de frecuencia por una impresión de la facilidad con que unos casos vienen a la mente. Por ejemplo, la percepción del riesgo de accidentes aéreos o de huracanes aumenta brevemente después de que se produzcan casos muy publicitados.

En teoría, un juicio sobre el riesgo debería basarse en un promedio a largo plazo. En realidad, se da más importancia a los incidentes recientes porque se los recuerda con más facilidad. La sustitución de la evaluación de la frecuencia por un juicio condicionado por la sencillez con que se recuerdan los ejemplos se conoce como «heurística de la disponibilidad».

La sustitución de un juicio difícil por otro fácil no se limita a estos ejemplos. De hecho, es muy común. La respuesta a una pregunta más fácil puede considerarse como un procedimiento general para responder a una pregunta que podría dejarnos mudos. Veamos cómo tendemos a responder a cada una de las siguientes preguntas utilizando el sustituto más fácil:

¿*Creo en el cambio climático*?
¿Me fío de las personas que dicen que existe?
¿*Creo que este cirujano es competente*?
¿Habla este hombre con seguridad y autoridad?
¿*Se completará el proyecto en el plazo previsto*?
¿Avanza en el plazo previsto?
¿*Es necesaria la energía nuclear*?
¿Me echa para atrás la palabra «nuclear»?
¿*Estoy satisfecho con mi vida en general*?
¿Cuál es mi estado de ánimo en este momento?

Independientemente de lo que se pregunte, la sustitución de una pregunta por otra conducirá a una respuesta que no otorga a los diferentes aspectos

de las pruebas su peso adecuado, y una ponderación incorrecta de las pruebas da lugar de forma inevitable a un error. Por ejemplo, una respuesta completa a una pregunta sobre la satisfacción con la vida requiere claramente consultar algo más que el estado de ánimo actual, pero parece claro que el estado de ánimo cobra un peso excesivo.

Del mismo modo, la sustitución de la probabilidad por la similitud hace que no se tengan en cuenta las tasas-base, que son irrelevantes cuando se juzga la similitud. Y factores como las variaciones irrelevantes en la estética del documento que presenta un plan de negocios deberían tener poco o ningún peso a la hora de evaluar el valor de una empresa. Cualquier repercusión que tengan en el juicio es probable que refleje una ponderación errónea de las pruebas, y producirá un error.

SESGOS DE CONCLUSIÓN

En un momento clave[5] del guion[6] de *El retorno del Jedi*, la tercera película de *La guerra de las galaxias*, George Lucas, el cerebro de la serie, mantuvo una acalorada discusión con Lawrence Kasdan, su gran colaborador. Este aconsejó lo siguiente a Lucas: «Creo que deberías matar a Luke y hacer que Leia tome el relevo». Lucas rechazó de inmediato la idea. Kasdan añadió que, si Luke vivía, otro personaje importante debía morir. Lucas volvió a discrepar y terminó diciendo: «No se va por ahí matando gente». Kasdan respondió con una sentida reivindicación de la naturaleza del cine. Le explicó a Lucas que «la película tiene más peso emocional si se pierde a alguien querido en el camino; el viaje tiene más impacto».

La respuesta de Lucas fue rápida e inequívoca: «No me gusta y no me lo creo».

El proceso del pensamiento parece aquí bastante diferente del que el lector experimentó cuando pensó en Bill, el contable que toca jazz. Volvamos a leer la respuesta de Lucas: «no gustarle» precede a «no creerlo». Lucas dio una respuesta automática a la sugerencia de Kasdan. Esa respuesta le ayudó a motivar su juicio (aunque resultara correcto).

Este ejemplo ilustra un tipo diferente de sesgo, que llamamos «sesgo de conclusión» o «prejuicio». Al igual que Lucas, a menudo iniciamos el proceso de juzgar con la inclinación a una conclusión particular. Cuando lo hacemos, dejamos que nuestro pensamiento rápido e intuitivo del Sistema 1

nos oriente a una conclusión. O bien llegamos a esa conclusión de forma precipitada y nos saltamos el proceso de recopilación e integración de información, o bien movilizamos el pensamiento del Sistema 2 —el pensamiento deliberativo— para elaborar argumentos que apoyen nuestro prejuicio. En este caso, las pruebas serán selectivas y distorsionadas: debido al «sesgo de confirmación» y al «sesgo de deseabilidad», tenderemos a reunir e interpretar las pruebas de forma selectiva para preferir un juicio que ya creemos verdadero o que deseamos que lo sea.[7]

La gente suele pergeñar racionalizaciones plausibles para sus juicios y, de hecho, pensará que estas son la causa de sus creencias. Una buena prueba del papel de los prejuicios es imaginar que los argumentos que aparentemente sustentan nuestra creencia se demuestran de forma repentina como nulos. Kasdan, por ejemplo, podría haber señalado a Lucas que «No se va por ahí matando gente» no es un argumento convincente. El autor de *Romeo y Julieta* no habría estado de acuerdo con Lucas, y si los guionistas de *Los Soprano* y *Juego de tronos* hubieran estado en contra de matar personajes, probablemente ambas series se habrían terminado en su primera temporada. Sin embargo, podemos apostar a que un fuerte contraargumento no habría hecho cambiar de opinión a Lucas. En su lugar, habría aportado otros argumentos para apoyar su juicio.(Por ejemplo, «*La guerra de las galaxias* es diferente».)

Los prejuicios son evidentes allá donde se manifiestan. Como la reacción de Lucas, tienen a menudo un componente emocional. El psicólogo Paul Slovic lo denomina «heurística del afecto»: las personas determinan lo que piensan consultando sus sentimientos. Nos gustan la mayoría de las cosas de los políticos que preferimos y nos disgustan incluso las miradas y las voces de los políticos que rechazamos. Esa es una de las razones por las que las empresas inteligentes se esfuerzan tanto en atribuir un efecto positivo a su marca. Los profesores suelen notar que, en un año en el que su estilo de enseñanza obtiene altas calificaciones, los estudiantes también dan una alta calificación al material del curso. En un año en el que a los estudiantes no les gusta tanto su enseñanza, dan una baja calificación a lecturas obligatorias idénticas. El mismo mecanismo funciona incluso cuando no hay emoción de por medio: con independencia de las verdaderas razones de nuestras creencias, nos inclinaremos a aceptar cualquier argumento que parezca respaldarlas, aunque el razonamiento sea erróneo.[8]

Un ejemplo más sutil de un sesgo de conclusión es el «efecto de anclaje», que es el efecto que un número arbitrario produce en personas que deben hacer un juicio cuantitativo. En una típica demostración, se le pueden presentar a esas personas varios artículos cuyo precio no es fácil de adivinar, como una botella de un vino que no conocen.[9] Se les pide que anoten las dos últimas cifras de su número de la Seguridad Social y que indiquen si pagarían esa cantidad por la botella. Finalmente, se les pide que indiquen la cantidad máxima que estarían dispuestas a pagar por ella. Los resultados muestran que el anclaje en el número de la Seguridad Social afecta al precio final. En un estudio, las personas cuyos números de la Seguridad Social marcaban un anclaje alto (más de ochenta dólares) se declararon dispuestas a pagar unas tres veces más que las que tenían un anclaje bajo (menos de veinte dólares).

Está claro que el número de la Seguridad Social no debería tener tan gran efecto en un juicio sobre cuánto vale una botella de vino, pero lo tiene. El anclaje es un efecto extremadamente fuerte y a menudo se utiliza de un modo deliberado en negociaciones.[10] Tanto si uno regatea en un bazar como si participa en una compleja negociación comercial, probablemente tenga ventaja si es el primero en hablar, porque el receptor del anclaje se ve incitado de forma involuntaria a pensar en las formas en que su oferta podría ser razonable. Las personas siempre intentan dar sentido a lo que oyen; cuando se encuentran con una cifra inverosímil, automáticamente abren su mente a consideraciones que reducirían su inverosimilitud.

EXCESIVA COHERENCIA

He aquí otro ejemplo que ayudará a experimentar un tercer tipo de sesgo. Se trata de leer la descripción de un candidato a un puesto ejecutivo. La descripción consiste en cuatro adjetivos, cada uno de ellos escrito en una carta. Las cartas se barajan, y las dos primeras tienen estos dos descriptores:

Inteligente, Perseverante.

Sería razonable aplazar el juicio hasta que la información sea completa, pero no es esto lo que ocurre: ya se tiene una evaluación del candidato, y es

positiva. Este juicio simplemente se produce. No se tenía ningún control sobre el proceso, y aplazar el juicio no era una opción.

A continuación, se leen las dos últimas cartas. En ellas está la descripción completa:

Inteligente, Perseverante, Astuto, Sin principios.

La evaluación ya no es favorable, pero no cambia lo suficiente. A modo de comparación, considérese la siguiente descripción, que otra barajadura podría haber producido:

Sin principios, Astuto, Perseverante, Inteligente.

Esta segunda descripción emplea los mismos adjetivos y, sin embargo, debido al orden en que se introducen, es con claridad mucho menos atrayente que la primera. La palabra «astuto» solo era un poco negativa cuando seguía a «inteligente» y «perseverante», porque se seguía creyendo (sin razón) que las intenciones del ejecutivo eran buenas. Sin embargo, cuando sigue a «sin principios», la palabra «astuto» es horrible. En este contexto, la perseverancia y la inteligencia ya no son positivas: hacen que una mala persona sea aún más peligrosa.

Este experimento ilustra una «coherencia excesiva»: nos formamos impresiones coherentes con rapidez y somos lentos en cambiarlas.[11] En este ejemplo, adoptamos de inmediato una actitud positiva hacia el candidato a la luz de unas pocas pruebas. El sesgo de confirmación —la misma tendencia que nos lleva, cuando tenemos un prejuicio, a ignorar por completo las pruebas que lo contradicen— nos hace asignar menos importancia de la que deberíamos a datos ulteriores. (Otro término para describir este fenómeno es el «efecto halo», porque el candidato fue evaluado por el «halo» positivo de la primera impresión. En el capítulo 24 veremos que el «efecto halo» es un serio problema en las decisiones de contratación.)

He aquí otro ejemplo. En Estados Unidos, los funcionarios públicos han exigido a las cadenas de restaurantes que incluyan etiquetas de calorías para garantizar que los consumidores vean las calorías asociadas a, por ejemplo, las hamburguesas con queso, las hamburguesas solas y las ensaladas. Después de ver esas etiquetas, ¿cambian los consumidores sus elecciones?

Las pruebas son controvertidas y ambivalentes. Sin embargo, en un estudio revelador se descubrió que los consumidores eran más propensos a verse afectados por las etiquetas de calorías si estaban colocadas a la izquierda del alimento en lugar de a la derecha.[12] Cuando las calorías figuran a la izquierda, los consumidores reciben primero esa información, y evidentemente antes de ver la comida piensan: «¡Tiene muchas calorías!» o «¡No tiene tantas calorías!». Su reacción inicial, positiva o negativa, afecta en gran medida a sus elecciones. En cambio, cuando la gente ve primero el alimento, aparentemente piensa: «¡Delicioso!» o «¡No es tan bueno!» antes de ver la etiqueta de calorías. También en este caso, su reacción inicial afecta en gran medida a sus elecciones. Esta hipótesis se ve respaldada por el descubrimiento que hicieron los autores de que, para los hablantes de hebreo, que leen de derecha a izquierda, la etiqueta de calorías tiene una repercusión significativamente mayor si está a la derecha en lugar de a la izquierda.

En general, sacamos conclusiones precipitadas y nos atenemos a ellas. Creemos que basamos nuestras opiniones en la evidencia, pero es probable que la evidencia que consideramos y nuestra interpretación de esta estén distorsionadas, al menos hasta cierto punto, para que se ajusten a nuestro precipitado juicio inicial. Como resultado, mantenemos la coherencia de toda la historia que se ha instalado en nuestra mente. Este proceso está bien si las conclusiones son correctas, pero, cuando la evaluación inicial es errónea, la tendencia de atenernos a ella frente a pruebas que la contradicen puede amplificar los errores. Y este efecto es difícil de controlar, porque la información que hemos visto u oído es imposible de ignorar y, a menudo, difícil de olvidar. En los tribunales, los jueces a veces instruyen a los miembros del jurado para que no tengan en cuenta una prueba improcedente que han escuchado, pero esta instrucción no es realista (aunque puede ser útil en las deliberaciones del jurado, donde los argumentos basados explícitamente en esa prueba pueden ser rechazados).

LOS SESGOS PSICOLÓGICOS CAUSAN RUIDO

Hemos presentado brevemente tres tipos de sesgos que operan de formas distintas: los sesgos de sustitución, que llevan a una ponderación errónea de las pruebas; los sesgos de conclusión, que nos llevan a obviar las pruebas o

a considerarlas de una forma distorsionada, y la coherencia excesiva, que magnifica el efecto de las impresiones iniciales y reduce el efecto de la información que las contradice. Los tres tipos de sesgo pueden, por supuesto, producir un sesgo estadístico. También pueden producir ruido.

Empecemos por la sustitución. Mayoritariamente se juzga la probabilidad de que Bill sea contable por la similitud de su perfil con un estereotipo: el resultado, en este experimento, es un sesgo compartido. Si todos los participantes cometen el mismo error, no hay ruido. Sin embargo, la sustitución no siempre produce esa unanimidad. Cuando la pregunta «¿Existe el cambio climático?» se sustituye por «¿Me fío de las personas que dicen que existe?», es fácil ver que la respuesta variará de una persona a otra en función de sus círculos sociales, de sus fuentes de información preferidas, de su afiliación política, etcétera. El mismo sesgo psicológico crea juicios variables y ruido entre personas.

La sustitución también puede ser una fuente de ruido de ocasión. Si una pregunta por la satisfacción con la vida se responde consultando el estado de ánimo inmediato, la respuesta variará de forma inevitable para la misma persona de un momento a otro. Una mañana feliz puede ir seguida de una tarde angustiosa, y los cambios de humor a lo largo del tiempo pueden dar lugar a afirmaciones sobre la satisfacción muy diferentes según el momento en que llame el entrevistador. En el capítulo 7 repasamos ejemplos de ruido de ocasión que puede ser atribuido a sesgos psicológicos.

También los prejuicios producen sesgo y ruido. Volvamos a un ejemplo que mencionamos en la introducción: las escandalosas disparidades en el porcentaje de solicitantes de asilo que los jueces aceptan. Cuando un juez acepta al 5 por ciento de los solicitantes y otro en el mismo juzgado acepta al 88 por ciento, podemos estar bastante seguros de que están mentalmente sesgados en diferentes direcciones. Desde una perspectiva más amplia, las diferencias individuales en los sesgos pueden causar un enorme ruido en el sistema. Por supuesto, este último también puede estar sesgado en la medida en que todos (o la mayoría de) los jueces manifiesten un sesgo similar.

Por último, la coherencia excesiva puede producir sesgo o ruido en función de si la secuencia de información y el significado que se le da son idénticos para todos (o la mayoría de) los juzgadores. Pensemos, por ejemplo, en un candidato físicamente atractivo cuya buena presencia crea

una primera impresión positiva en la mayoría de los contratadores. Si el aspecto físico es irrelevante en el puesto para el que se considera al candidato, este halo positivo dará lugar a un error compartido: un sesgo.

Por otra parte, muchas decisiones complejas requieren recopilar información que llega en un orden esencialmente aleatorio. Pensemos en los tasadores de reclamaciones del capítulo 2. El orden en que se dispone de los datos de una reclamación varía de forma aleatoria de un tasador a otro y de un caso a otro, lo que provoca una variación aleatoria de las impresiones iniciales. Un exceso de coherencia significa que estas variaciones aleatorias producirán distorsiones aleatorias en los juicios finales. El efecto será ruido del sistema.

En resumen, los sesgos psicológicos son, como mecanismo, universales, y, a menudo, producen errores compartidos. Sin embargo, cuando hay grandes diferencias individuales en los sesgos (diferentes prejuicios), o cuando el efecto de los sesgos depende del contexto (diferentes desencadenantes), habrá ruido.

Tanto el sesgo como el ruido generan errores, lo que indica que todo lo que reduzca los sesgos psicológicos mejorará el juicio. Volveremos sobre el tema de la eliminación del sesgo en la quinta parte. Por ahora, proseguiremos con nuestro análisis del proceso del juicio.

SOBRE HEURÍSTICAS, SESGOS Y RUIDO

- «Sabemos que tenemos sesgos psicológicos, pero debemos resistir el impulso de culpar de cada error a “sesgos” no especificados.»
- «Cuando sustituimos la pregunta a la que debemos responder por otra más fácil, es inevitable que se produzcan errores. Por ejemplo, ignoramos la tasa-base cuando juzgamos la probabilidad por la similitud.»
- «Los prejuicios y otros sesgos de conclusión llevan a las personas a distorsionar las pruebas en favor de su posición inicial.»
- «Nos formamos ideas rápidamente, y nos aferramos a ellas incluso cuando nos llega información que las contradice. Esta tendencia se llama “coherencia

excesiva”.»

- «Los sesgos psicológicos causan sesgos estadísticos si muchas personas comparten los mismos sesgos. Sin embargo, en muchos casos no los comparten. En esos casos, los sesgos psicológicos crean ruido del sistema.»

14
La operación de emparejamiento

Observemos el cielo. ¿Qué probabilidad hay de que llueva en dos horas?

Seguramente no nos costará responder a esta pregunta. El juicio que hagamos —por ejemplo, es «muy probable» que llueva pronto— no nos exigirá ningún esfuerzo. De alguna manera, nuestra evaluación de la nubosidad se convierte en un juicio de probabilidad.

Este es un ejemplo elemental de «emparejamiento». Hemos descrito el «juicio» como una operación que asigna un valor de una escala a una impresión subjetiva (o a un aspecto de una impresión). El emparejamiento es una parte esencial de esa operación. Cuando respondemos a la pregunta: «En una escala del 1 al 10, ¿cuál es su estado de ánimo?», o se nos pide que «Por favor, ponga de una a cinco estrellas a su experiencia con la compra de esta mañana», estamos emparejando: la tarea es dar un valor de una escala que se ajuste a un estado de ánimo o a una experiencia.

EMPAREJAMIENTO Y COHERENCIA

Ya analizamos el ejemplo de Bill en el capítulo anterior, y aquí lo tenemos de nuevo: «Bill tiene treinta y tres años. Es inteligente, pero poco imaginativo, compulsivo y, en general, apático. En la escuela era fuerte en matemáticas, pero flojo en ciencias sociales y humanidades». Pedimos al lector que estimara la probabilidad de que Bill tuviera varias ocupaciones y aficiones, y vimos que respondió a esta pregunta sustituyendo un juicio de probabilidad por otro de similitud. En realidad, no se había preguntado por la probabilidad de que Bill fuese contable, sino por su similitud con el estereotipo de esa profesión. Ahora pasamos a una pregunta que dejamos sin respuesta: cómo hizo ese juicio de similitud.

No es difícil evaluar hasta qué punto la descripción de Bill coincide con estereotipos de profesiones y aficiones. Está claro que Bill se parece menos a un típico músico de jazz que a un contable, y aún menos a un surfista. El ejemplo ilustra la extraordinaria versatilidad del emparejamiento, que es

muy evidente en los juicios sobre personas. Apenas hay un límite a las preguntas que el lector podría haber respondido sobre Bill. Por ejemplo: ¿cómo se sentiría si se quedara abandonado en una isla desierta con él? Seguramente, a partir de la escasa información facilitada, la respuesta a esta pregunta cualquiera la intuye de manera inmediata. Sin embargo, puede llegar una nueva noticia: Bill, tal y como el lector lo conoce, resulta ser un explorador curtido con extraordinarias habilidades de supervivencia. Si esto llega a sorprenderlo (y probablemente lo haga), percibirá un fallo en la coherencia inicial.

La sorpresa es grande, porque la nueva información es incompatible con la imagen que antes tenía de Bill. Imaginemos ahora que las capacidades y habilidades de Bill para la supervivencia se hubieran incluido en la descripción original. El lector habría acabado formándose una idea general diferente del hombre, quizá la de alguien que cobra vida solo en entornos naturales. La idea general sobre Bill habría sido menos coherente y, por tanto, más difícil de adecuar a categorías de profesiones o aficiones, pero habría percibido mucha menos discordancia que antes.

Las pistas contradictorias hacen más difícil tener una sensación de coherencia y encontrar un juicio que sea satisfactorio. La presencia de pistas conflictivas caracteriza a los juicios complejos, en los que esperamos encontrar mucho ruido. El problema de Gambardi, en el que algunos de los indicios eran positivos y otros negativos, fue un juicio de este tipo. Volveremos sobre estos juicios complejos en el capítulo 16. En el resto de este capítulo nos centraremos en juicios relativamente sencillos (especialmente en los que emplean «escalas de intensidad»).

EMPAREJAMIENTO DE INTENSIDADES

Algunas de las escalas que empleamos para expresar juicios son cualitativas: las profesiones, las aficiones y los diagnósticos médicos serían ejemplos de ello. Se identifican por el hecho de que los valores de la escala no están ordenados: el rojo no es más ni menos que el azul.

Sin embargo, muchos juicios emplean escalas de intensidad cuantitativa. Las mediciones físicas de tamaño, peso, brillo, temperatura o intensidad sonora, así como las medidas de coste o valor y los juicios de probabilidad o frecuencia, son estimaciones cuantitativas. También lo son los juicios

expresados en escalas más abstractas, como la confianza, la fuerza, el atractivo, la ira, el miedo, la inmoralidad o la severidad de las sentencias judiciales.

El rasgo distintivo que comparten estas magnitudes medibles cuantitativas es que la pregunta «¿Qué es más?» se puede responder para cualquier par de valores en la misma magnitud. Podemos decir que una flagelación es un castigo más severo que un tirón de orejas, o que *Hamlet* nos gusta más que *Esperando a Godot*, igual que podemos decir que el sol es más brillante que la luna, que un elefante pesa más que un hámster y que la temperatura media en Miami es más alta que en Toronto.

Las personas tienen una notable capacidad intuitiva para emparejar intensidades entre magnitudes no relacionadas mediante la asignación de una escala de intensidad a otra.[1] Podemos emparejar la intensidad de nuestro gusto por diferentes cantantes con la altura de los edificios de una ciudad. (Quien crea que Bob Dylan es particularmente grande, por ejemplo, podría emparejar su nivel de entusiasmo por él con el edificio más alto de su ciudad.) Alguien podría emparejar el nivel actual de discordia política en su país con la temperatura veraniega de una ciudad que conozca bien.(Si hay una notable armonía política, la podría equiparar a un día de verano con brisa de 70 grados centígrados en Nueva York.) Y si le pidieran que expresara su aprecio por un restaurante comparándolo con la extensión de una novela, en lugar de la habitual escala de valoración de 1 a 5 estrellas, esta petición le parecería bastante extraña, pero en absoluto inviable. (Su restaurante favorito podría ser como *Guerra y paz*.) En cada caso, es —extrañamente— bastante claro lo que se quiere decir.

En una conversación corriente, el rango de valores de una escala es una función del contexto. El comentario «Ha estado ahorrando mucho dinero» tiene un significado diferente cuando se está brindando por la jubilación de un banquero de inversiones que ha tenido mucho éxito que cuando se felicita a una adolescente que ha estado trabajando de canguro. Y el significado de palabras como «grande» y «pequeño» depende del todo de un marco de referencia. Podemos, por ejemplo, dar sentido a una afirmación como «El ratón grande corrió por la trompa del pequeño elefante».

El siguiente acertijo ilustra tanto el poder del emparejamiento como un error de juicio sistemático asociado a él.[2]

> Julie es una estudiante que se gradúa en una universidad. Lea la siguiente información sobre ella, y luego adivine su nota media (en la escala estándar de 0,0 a 4,0):
>
> Julie leía con fluidez cuando tenía cuatro años.
>
> ¿*Cuál es su nota media*?

Si el lector está familiarizado con el sistema de puntuación de Estados Unidos, le vendrá enseguida a la mente una cifra que probablemente se acerque a 3,7 o 3,8. El hecho de que una estimación de la nota media de Julie se le ocurriera al instante, ilustra el proceso de emparejamiento que acabamos de describir.

Al principio, habrá evaluado lo precoz que era Julie como lectora. La evaluación fue fácil porque Julie leía de forma atípicamente precoz, y esa precocidad situaba a Julie en una categoría de alguna escala. Si tuviera que describir la escala que utilizó, probablemente diría que su categoría más alta es algo así como «extraordinariamente precoz», y señalaría que Julie no pertenece del todo a esa categoría (algunos niños leen antes de los dos años). Julie probablemente pertenezca a la siguiente, al grupo de niños «atípicamente, pero no extraordinariamente, precoces».

A continuación, habrá emparejado un juicio sobre la nota media con su evaluación de Julie. Aunque no era consciente de ello, debió de buscar un valor de la nota media que también encajara en la etiqueta «atípico, pero no extraordinario». Una «predicción de emparejamiento» vino a su mente, al parecer de la nada, cuando conoció la historia de Julie.

Llevar a cabo de forma deliberada los cálculos necesarios para realizar estas tareas de evaluación y emparejamiento requeriría bastante tiempo, pero, en el pensamiento rápido del Sistema 1, el juicio se hace con rapidez y sin esfuerzo. La historia que contamos aquí sobre la adivinación de la nota media de Julie implica una compleja secuencia de eventos mentales en varias etapas que no puede observarse directamente. La especificidad del mecanismo mental del emparejamiento es poco frecuente en psicología, pero las pruebas que lo demuestran son excepcionalmente concluyentes. Podemos estar seguros, por muchos experimentos similares, de que las dos

preguntas siguientes, cuando se hacen a distintos grupos de personas, darán exactamente las mismas cifras:[3]

• ¿Qué porcentaje de la clase de Julie lee a una edad más temprana que ella?
• ¿Qué porcentaje de la clase de Julie tiene una nota media más alta que ella?

La primera pregunta es fácil de responder: simplemente nos pide evaluar las pruebas que nos dieron sobre Julie. La segunda pregunta, que requiere una predicción distante, es, desde luego, más difícil, pero es intuitivamente tentador contestarla respondiendo a la primera.

Las dos preguntas que nos hacemos sobre Julie son análogas a dos preguntas que describimos como universalmente confusas en un debate anterior sobre la ilusión de validez. La primera pregunta sobre Julie requiere que evaluemos la «intensidad» de la información que tenemos sobre su caso. La segunda pregunta se refiere a la intensidad de una predicción. Y sospechamos que sigue siendo difícil distinguirla.

La predicción intuitiva de la nota media de Julie es un caso del mecanismo psicológico que describimos en el capítulo 13: la sustitución de una pregunta difícil por otra fácil. El Sistema 1 simplifica una pregunta de predicción difícil respondiendo a otra más fácil: ¿cuánto le impresiona Julie como lectora de cuatro años? Para pasar luego directamente de la edad de lectura, medida en años, a la nota media, medida en puntos, se requiere un paso adicional de emparejamiento.

La sustitución se produce, por supuesto, solo si la información disponible es relevante. Si todo lo que supiéramos de Julie fuera que es una corredora veloz o una bailarina mediocre, no tendríamos ninguna información. Sin embargo, cualquier hecho que pueda interpretarse como un indicio plausible de inteligencia es, con toda probabilidad, un sustituto aceptable.

Inevitablemente, sustituir una pregunta por otra provoca errores cuando las verdaderas respuestas a las dos preguntas son distintas. Sustituir la edad de lectura por el promedio de notas, aunque parezca plausible, es manifiestamente absurdo. Para ver por qué, pensemos en lo que podría haberle acontecido a Julie desde que tenía cuatro años. Podría haber tenido un terrible accidente. Sus padres podrían haberse divorciado de forma traumática. Podría haber encontrado un profesor motivador que ejerciera gran influencia sobre ella. Podría haberse quedado embarazada. Cualquiera

de estos acontecimientos, y muchos más, podrían haber afectado a sus estudios en la universidad.

La predicción de emparejamiento solo puede justificarse si la precocidad lectora y la nota media universitaria están del todo correlacionadas, lo que claramente no es el caso. Por otro lado, ignorar por completo la información sobre la edad de lectura de Julie también sería un error, ya que su edad de lectura proporciona cierta información relevante. La predicción óptima debe situarse entre estos dos extremos de conocimiento completo y conocimiento cero.

¿Qué sabemos de un caso cuando nada concreto sabemos de él? ¿Solo la categoría a que pertenece? La respuesta a esta pregunta es lo que hemos llamado la «visión externa» del caso. Si se nos pidiera que predijéramos la nota media de Julie, pero no se nos diera ninguna información sobre ella, seguramente predeciríamos la media general (quizá 3,2). Esta es la predicción de visión externa. Por tanto, la mejor estimación de la nota media de Julie debe ser superior a 3,2 e inferior a 3,8. La ubicación precisa de la estimación depende del valor predictivo de la información: cuanto más se confíe en la edad de lectura como predictor de la nota media, tanto más alta será la estimación. En el caso de Julie, la información es desde luego bastante escasa, y la predicción más razonable será, en consecuencia, la más cercana a la media general. Hay una forma técnica, pero bastante fácil, de corregir el error de las predicciones de emparejamiento (la detallamos en el apéndice C).

Aunque conduzcan a predicciones estadísticamente absurdas, es difícil resistirse a las predicciones que se emparejan con las evidencias. Los directores de ventas suelen dar por sentado que el vendedor que tuvo más éxito que el resto del equipo de ventas el último año seguirá teniéndolo. Los altos ejecutivos a veces conocen a un candidato con un talento excepcional y se imaginan cómo el nuevo contratado llegará a la cima de la organización. Los productores suelen anticipar que la siguiente película de un director cuya película anterior fue un éxito tendrá el mismo éxito.

Lo más probable es que estos ejemplos de predicciones emparejadas acaben en decepción. Por otra parte, las predicciones emparejadas que se hacen cuando las cosas están en su peor momento tienen más probabilidades de ser excesivamente negativas. Las predicciones intuitivas que se emparejan con las evidencias son demasiado extremas, tanto si son

optimistas como si son pesimistas. (El término técnico para estos errores de predicción es que son «no regresivos», porque no tienen en cuenta un fenómeno estadístico llamado «regresión a la media».)

Sin embargo, hay que tener en cuenta que la sustitución y el emparejamiento no siempre rigen las predicciones. En el lenguaje de los dos sistemas, el intuitivo Sistema 1 propone soluciones asociativas rápidas a los problemas a medida que surgen, pero estas intuiciones deben ser refrendadas por el más reflexivo Sistema 2 antes de convertirse en creencias. A veces se rechazan las predicciones de emparejamiento en favor de respuestas más complejas. Por ejemplo, las personas son más reacias a emparejar las predicciones con las evidencias desfavorables que con las favorables. Sospechamos que, en el caso de Julie, el lector habría dudado en hacer una predicción emparejada de un rendimiento universitario inferior si Julie hubiera sido una lectora tardía. La asimetría entre las predicciones favorables y desfavorables desaparece cuando se dispone de más información.

Ofrecemos la visión externa como un correctivo para las predicciones intuitivas de todo tipo. En un debate anterior sobre las perspectivas de futuro de Michael Gambardi, por ejemplo, recomendamos anclar el juicio sobre la probabilidad de éxito de Michael en la tasa-base pertinente (la tasa de éxito de dos años para los directores ejecutivos entrantes). En el caso de las predicciones cuantitativas, como la nota media de Julie, adoptar la visión externa significa anclar la predicción en el resultado medio. La visión externa solo se puede descuidar en problemas muy fáciles, cuando la información disponible apoya una predicción que se puede hacer con total confianza. Cuando es necesario un juicio serio, la visión externa debe ser parte de la solución.

RUIDO EN EL EMPAREJAMIENTO: LIMITACIONES DEL JUICIO ABSOLUTO

Nuestra limitada capacidad para distinguir categorías en escalas de intensidad restringe la exactitud de la operación de emparejamiento. Palabras como «grande» o «rico» asignan la misma etiqueta a un rango de valores en la dimensión del tamaño o la riqueza. Esta es una fuente de ruido potencialmente importante.

El banquero de inversiones que se jubila merece sin duda la etiqueta de «rico», pero ¿hasta qué punto lo es? Tenemos muchos adjetivos para elegir: «adinerado», «acaudalado», «opulento», «acomodado» y otros. Si nos dieran descripciones detalladas de la riqueza de algunos individuos y tuviéramos que emplear un adjetivo para cada uno, ¿cuántas categorías distintas podríamos establecer sin recurrir a comparaciones detalladas entre casos?

El número de categorías que podemos distinguir en una escala de intensidad aparece en el título de un artículo clásico de la psicología publicado en 1956: «The Magical Number Seven, Plus or Minus Two».[4] Fuera de este límite, la gente tiende a empezar a cometer errores (por ejemplo, asignar a A una categoría superior a B, cuando, de hecho, calificaría a B más alto que a A en una comparación efectiva).

Imaginemos un conjunto de líneas de cuatro longitudes diferentes entre 5 y 10 centímetros, con cada línea igual de larga que la siguiente. Se nos muestra una línea cada vez, y tenemos que decir un número entre 1 y 4, donde 1 corresponde a la línea más corta y 4 a la más larga. La tarea es fácil. Ahora supongamos que se nos muestran líneas de cinco longitudes diferentes y que tenemos que repetir la tarea diciendo números del 1 al 5. Sigue siendo fácil. ¿Cuándo empezaremos a cometer errores? Alrededor del número mágico de siete líneas. Sorprendentemente, esta cifra depende muy poco del rango de longitudes de las líneas: si las líneas estuvieran espaciadas entre 5 y 15 centímetros, en lugar de entre 5 y 10, seguiríamos cometiendo errores más allá de siete líneas. El mismo resultado se obtiene cuando se presentan tonos que varían en intensidad, o luces de diferente brillo. Existe un límite real en la capacidad de las personas para poner etiquetas distintas a estímulos en una dimensión, y ese límite está en torno a siete etiquetas.

Este límite de nuestro poder de diferenciación es importante, porque nuestra capacidad de emparejar valores en dimensiones de intensidad no puede ser mayor que nuestra capacidad de asignar valores en esas dimensiones. La operación de emparejamiento es una herramienta versátil del pensamiento rápido del Sistema 1 y el núcleo de muchos juicios intuitivos, pero es tosca.

El número mágico no es una restricción absoluta. Se puede enseñar a las personas a hacer distinciones más detalladas mediante la categorización

jerárquica. Por ejemplo, podemos sin duda distinguir varias categorías de riqueza entre los multimillonarios, y los jueces pueden distinguir los grados de severidad en múltiples categorías de delitos, ordenados a su vez por gravedad. Sin embargo, para que este proceso de perfeccionamiento funcione, las categorías deben existir previamente, y sus límites han de ser claros. Al asignar etiquetas a un conjunto de líneas, no podemos decidir separar las líneas más largas de las más cortas y tratarlas como dos categorías distintas. La categorización no está bajo control voluntario cuando se está en el modo de pensamiento rápido.

Hay una forma de superar la limitada resolución de las escalas adjetivas: en lugar de utilizar etiquetas, usar comparaciones. Nuestra capacidad para comparar casos es mucho mejor que nuestra capacidad para situarlos en una escala.

Pensemos en lo que haríamos si se nos pidiera que utilizáramos una escala de calidad de veinte puntos para evaluar un gran conjunto de restaurantes o de cantantes. Una escala de cinco estrellas sería fácilmente manejable, pero no se podría mantener una fiabilidad completa con una escala de veinte puntos. (Joe’s Pizza merece tres estrellas, pero ¿un once o un doce?) La solución a este problema es sencilla, aunque requiere tiempo. En primer lugar, se calificaría a los restaurantes, o cantantes, utilizando la escala de valoración de cinco puntos para clasificarlos en cinco categorías. A continuación, se clasificarán los casos dentro de cada categoría, lo que normalmente se podrá hacer con solo algunos empates. De ese modo, probablemente sepamos si preferimos Joe’s Pizza a Fred’s Burgers, o Taylor Swift a Bob Dylan, aunque los hayamos colocado en la misma categoría. Para simplificar las cosas, ahora podríamos distinguir cuatro niveles dentro de cada una de las cinco categorías. Probablemente podamos diferenciar los niveles de desdén incluso entre los cantantes que menos nos gustan.

La psicología de este ejercicio es sencilla. Las comparaciones explícitas entre objetos de juicio admiten diferenciaciones mucho más detalladas que las estimaciones de objetos evaluados de uno en uno. Los juicios sobre la longitud de las líneas cuentan una historia similar: nuestra capacidad para comparar la longitud de las líneas que se muestran en sucesión inmediata es mucho mejor que nuestra capacidad para etiquetar longitudes, y será aún más precisa si comparamos líneas que están a la vista al mismo tiempo.

La ventaja de los juicios comparativos se aplica a muchos campos. Si tenemos una idea aproximada de la riqueza de las personas, lo haremos mejor comparando individuos del mismo rango que etiquetando su riqueza individualmente. Si calificamos las redacciones, seremos más precisos si las clasificamos de mejor a peor que leyéndolas y calificándolas una tras otra. Los juicios comparativos o relativos son más sensibles que los categóricos o absolutos. Como sugieren estos ejemplos, también requieren más esfuerzo y más tiempo.

Clasificar los objetos de forma individual en escalas que son explícitamente comparativas conserva algunas de las ventajas del juicio comparativo. En algunos contextos, sobre todo en la educación, las recomendaciones de candidatos para su aceptación o promoción suelen requerir que el recomendador sitúe al candidato en el «5 por ciento superior» o el «20 por ciento superior» de alguna población designada, como «los estudiantes a los que hemos dado clase» o «los programadores con el mismo nivel de experiencia». Estas calificaciones rara vez merecen ser tomadas al pie de la letra, porque no hay manera de responsabilizar a los recomendadores de utilizar la escala de manera correcta. La responsabilidad es posible en algunos contextos: cuando los directivos califican a los empleados o cuando los analistas evalúan las inversiones, una persona que asigna el 90 por ciento de los casos a la categoría del «20 por ciento superior» puede ser identificada y corregida. El uso de juicios comparativos es uno de los remedios contra el ruido que analizaremos en la quinta parte.

Muchas tareas de enjuiciamiento requieren que los casos individuales entren en una categoría de una escala (por ejemplo, una escala de acuerdo de siete puntos) o que se utilice un conjunto ordenado de adjetivos (por ejemplo, «improbable» o «extremadamente improbable» para determinar las probabilidades de acontecimientos). Este tipo de comparación es ruidosa porque es tosca. Los individuos pueden diferir en la interpretación de las etiquetas incluso cuando están de acuerdo en lo sustancial del juicio. Un procedimiento que obligue a realizar juicios explícitamente comparativos es probable que reduzca el ruido.[5] En el próximo capítulo, analizaremos más a fondo cómo el uso de escalas equivocadas puede aumentar el ruido.

SOBRE EL EMPAREJAMIENTO

- «Los dos decimos que esta película es muy buena, pero parece que tú la disfrutaste mucho menos que yo. Estamos usando las mismas palabras, pero ¿estamos empleando la misma escala?»
- «Pensamos que la segunda temporada de esta serie sería tan espectacular como la primera. Hicimos una predicción coincidente, y fue errónea.»
- «Es difícil mantener la coherencia a la hora de calificar estas redacciones. ¿Debería intentar clasificarlas en su lugar?»

15
Escalas

Imagine el lector que es miembro de un jurado en un juicio civil. Ha escuchado las pruebas, que se resumen a continuación, y deberá emitir algunos juicios sobre ellas.

JOAN GLOVER CONTRA GENERAL ASSISTANCE

Joan Glover, una niña de seis años, ingirió un gran número de pastillas de Allerfree, un medicamento para la alergia de venta libre, y necesitó una larga estancia en el hospital. Como la sobredosis debilitó su sistema respiratorio, será más susceptible de padecer enfermedades respiratorias, como asma y enfisema, durante el resto de su vida. El frasco de Allerfree utilizaba un tapón de seguridad a prueba de niños, pero de diseño inadecuado.

El fabricante de Allerfree es General Assistance, una gran empresa (con beneficios anuales de entre cien y doscientos millones de dólares) que produce una variedad de medicamentos sin receta. Una normativa federal exige que todos los frascos de medicamentos lleven tapones de seguridad a prueba de niños. General Assistance ha ignorado sistemáticamente el propósito de esta normativa al utilizar un tapón de seguridad a prueba de niños que tenía una tasa de fallos mucho más alta que la de otros del sector. Un documento interno de la empresa dice que «esta estúpida e innecesaria regulación federal nos hace desperdiciar dinero», y afirma que el riesgo de ser sancionado es bajo. El documento añade que, en cualquier caso, «las penas por infringir la normativa son extremadamente suaves; básicamente se nos pediría que mejoráramos los tapones de seguridad en el futuro». Aunque un funcionario de la Agencia de Alimentos y Medicamentos de Estados Unidos advirtió a la empresa sobre el defecto de su tapón de seguridad, la empresa decidió no tomar ninguna medida correctiva.

Ahora pedimos al lector que haga tres juicios y que no se precipite al elegir sus respuestas.

Indignación:
¿Cuál de las siguientes opciones expresa mejor su opinión sobre las acciones del demandado? (Marque con un círculo su respuesta.)

Completamente aceptable	Censurable	Escandaloso		Escandaloso		Absolutamente indignante
0	1	2	3	4	5	6

Intención punitiva:
Además de indemnizar por daños y perjuicios, ¿qué pena debe imponerse al demandado? (Marque con un círculo el número que mejor exprese su opinión sobre el nivel adecuado de la pena.)

Ninguna pena		Pena leve		Pena severa		Pena muy severa
0	1	2	3	4	5	6

Daños y perjuicios:
Además de las indemnizaciones por daños y perjuicios, ¿qué cuantía debe pagar la empresa demandada por daños «punibles» (si los hay) para disuadirla a ella y a otras de acciones similares en el futuro? (Escriba su respuesta en el siguiente espacio en blanco.)

..... €

El caso de Joan Glover es una versión ligeramente abreviada de otro utilizado en un estudio que dos de nosotros (Kahneman y Sunstein, junto con nuestro amigo y colaborador David Schkade) presentamos en 1998.[1] Describimos este estudio con cierto detalle en este capítulo porque queríamos que el lector experimentase una de las tareas que incluye el estudio, pues ahora la vemos como un ejemplo instructivo de una auditoría del ruido que insiste en muchos de los temas de este libro.

Este capítulo se centra en el papel de la «escala de respuestas» como fuente omnipresente de ruido. Las personas pueden diferir en sus juicios, no porque no estén de acuerdo en el fondo, sino porque utilizan la escala de diferentes maneras. Si alguien calificara el rendimiento de un empleado, podría decir que, en una escala de 0 a 6, el rendimiento fue de 4, lo que, en

su opinión, es bastante bueno. Otra persona podría decir que, en la misma escala, el rendimiento del empleado fue de 3, lo que, en su opinión, es también bastante bueno. La ambigüedad en la confección de las escalas es un problema general. Se han llevado a cabo muchas investigaciones sobre las dificultades de comunicación derivadas de expresiones vagas como «fuera de toda duda razonable»,[2] «pruebas claras y convincentes», «rendimiento sobresaliente» e «improbable que suceda».[3] Los juicios que se expresan en estas expresiones son inevitablemente ruidosos, porque son interpretados de forma diferente por quienes las dicen y por quienes las escuchan.

En el estudio para el que se redactó el caso de Joan Glover, observamos los efectos de una escala ambigua en una situación en la que esta tiene serias consecuencias. El tema del estudio fue el ruido en las indemnizaciones punitivas impuestas por un jurado. Como se puede deducir de la tercera pregunta sobre el caso de Joan Glover, en Estados Unidos (y en algunos otros países) la ley permite a los jurados, en casos civiles, imponer indemnizaciones punitivas a un demandado cuyas acciones fueron particularmente perjudiciales. Las indemnizaciones punitivas son complementarias de las indemnizaciones compensatorias por daños y perjuicios, destinadas a resarcir a los perjudicados. Cuando, como en el ejemplo de Glover, un producto ha causado lesiones y los demandantes se han querellado con éxito contra la empresa, se les asigna una cuantía para pagar sus facturas médicas y cualquier pérdida salarial. Sin embargo, también pueden recibir una indemnización punitiva, cuyo objetivo es enviar una advertencia a la empresa demandada y a otras similares. El comportamiento de General Assistance en este caso fue obviamente reprobable; entra en el rango de acciones por las que un jurado podría razonablemente imponer indemnizaciones punitivas.

Una de las principales preocupaciones que causa la institución de las indemnizaciones punitivas ha sido su imprevisibilidad. La misma infracción puede ser penada con indemnizaciones que van desde las muy modestas hasta las cuantiosas. Utilizando la terminología de este libro, diríamos que el sistema es ruidoso. Las solicitudes de indemnizaciones punitivas se deniegan a menudo, e incluso, si se conceden, no suelen añadirse a las indemnizaciones por daños y perjuicios. Sin embargo, hay llamativas excepciones, y las elevadísimas cuantías que los jurados imponen se antojan

a veces sorprendentes y arbitrarias. Un ejemplo que se menciona a menudo es una indemnización punitiva de cuatro millones de dólares impuesta a un concesionario de automóviles por no revelar el hecho de que el nuevo BMW del demandante había sido repintado.[4]

En nuestro estudio sobre las indemnizaciones punitivas, se pidió a ochocientos noventa y nueve participantes que evaluaran el caso de Joan Glover y otros nueve casos similares, todos ellos con demandantes que habían sufrido algún daño y que habían demandado a la empresa supuestamente responsable. A diferencia del lector, los participantes solo respondieron a una de las tres preguntas (indignación, intención punitiva o indemnización en dólares) en los diez casos. Los participantes se dividieron a su vez en grupos más pequeños, cada uno de los cuales fue asignado a una versión de cada caso. Las diferentes versiones variaron el daño sufrido por el demandante y los ingresos de la empresa demandada. Había un total de veintiocho escenarios. Nuestros objetivos eran poner a prueba una teoría sobre la psicología de las indemnizaciones punitivas e investigar el papel de la escala monetaria (en este caso, los dólares) como principal fuente de ruido en esta institución jurídica.

LA HIPÓTESIS DE LA INDIGNACIÓN

Los filósofos y los juristas han debatido durante siglos cómo determinar una pena justa. Sin embargo, nuestra hipótesis era que la pregunta que los filósofos encuentran difícil es bastante fácil para la gente común, que simplifica la tarea sustituyendo la pregunta difícil por otra fácil. La pregunta fácil, que se responde de inmediato cuando se pregunta cuánto debe pagar General Assitance, es: «¿Hasta qué punto estoy indignado?». La intensidad de la pena corresponderá entonces a la intensidad de la indignación.

Para contrastar esta hipótesis de la indignación, pedimos a diferentes grupos de participantes que respondieran a la pregunta de intención punitiva o a la de indignación. A continuación, comparamos las valoraciones medias obtenidas en las dos preguntas para los veintiocho escenarios utilizados en el estudio. Como se esperaba de la idea de la sustitución, la correlación entre las valoraciones medias de la indignación y de la intención punitiva fue de un casi perfecto .98 (PC = 94 %). Esta correlación apoya la hipótesis

de la indignación: la emoción de la indignación es el principal determinante de la intención punitiva.[5]

La indignación es el principal motor de la intención punitiva, pero no el único. ¿Notó el lector en el caso de Joan algo que le llamara más la atención cuando puntuó la intención punitiva que cuando puntuó la indignación? Si le ocurrió, sospechamos que fue el daño sufrido. Se puede saber si un comportamiento es indignante sin conocer sus consecuencias; en este caso, el comportamiento de General Assistance fue seguramente indignante. En cambio, las intuiciones sobre la intención punitiva tienen un aspecto de castigo toscamente expresado en el principio del ojo por ojo. El afán de castigo explica que la ley y los jurados traten de forma diferente la tentativa de asesinato y el asesinato; un aspirante a asesino que tenga la suerte de no lograr su objetivo será castigado con menos severidad.

Para averiguar si el daño marca realmente la diferencia en la intención punitiva, pero no en la indignación, mostramos a diferentes grupos de encuestados las versiones de «daño grave» y «daño leve» en el caso de Joan Glover y de varios otros. La versión de daño grave es la que el lector vio. En la versión de daño leve, Joan «tuvo que pasar varios días en un hospital, y ahora está muy traumatizada y no quiere tomar ninguna clase de pastillas. Cuando sus padres intentan que tome incluso medicamentos buenos para la salud como vitaminas, aspirinas o remedios para el resfriado, llora descontroladamente y dice que tiene miedo». Esta versión describe una experiencia traumática para la niña, pero un nivel de daño mucho menor que el daño a largo plazo descrito en la primera versión. Como era de esperar, las valoraciones medias de la indignación fueron casi idénticas para la versión de daño grave (4,24) y la de daño leve (4,19). Solo el comportamiento del demandado cuenta para la indignación, no sus consecuencias. En cambio, las valoraciones de la intención punitiva fueron de una media de 4,93 para los daños graves y de 4,65 para los leves, una diferencia pequeña pero estadísticamente fiable. La mediana de las indemnizaciones monetarias fue de dos millones de dólares para la versión de daños graves y de un millón para la versión de daños leves. Se obtuvieron resultados similares en otros casos.

Estos resultados ponen de manifiesto una característica clave del proceso de juzgar: el sutil efecto de la tarea del juicio sobre la ponderación de diferentes aspectos de las pruebas. Los participantes que puntuaron la

intención punitiva y la indignación no eran conscientes de que estaban tomando una posición sobre la cuestión filosófica de si la justicia debe ser de castigo. Ni siquiera eran conscientes de asignar un peso a las distintas características del caso. No obstante, asignaron un peso casi nulo al daño al puntuar la indignación y un peso significativo al mismo factor al determinar la pena. Recordemos que los participantes solo vieron una versión del caso; su asignación de un castigo mayor al daño mayor no fue una comparación explícita. Fue el resultado de una operación automática de emparejamiento en las dos condiciones. Las respuestas de los participantes se basaron más en el pensamiento rápido que en el lento.

ESCALAS RUIDOSAS

El segundo objetivo del estudio era averiguar por qué las indemnizaciones punitivas son ruidosas. Nuestra hipótesis era que los miembros de los jurados suelen estar de acuerdo en la severidad con la que desean que se castigue al demandado, pero difieren mucho en la forma de trasladar su intención punitiva a la escala de dólares.

El diseño del estudio permite comparar la cantidad de ruido en los juicios de los mismos casos en tres escalas: indignación, intención punitiva e indemnizaciones en dólares. Para medir el ruido, aplicamos el método que se utilizó para analizar los resultados de la auditoría del ruido de los jueces federales en el capítulo 6. Suponemos, como hicimos en ese análisis, que el promedio de juicios individuales de un caso puede tratarse como un valor no sesgado y justo. (Se trata de una suposición a efectos del análisis; insistimos en que podría ser errónea.) En un mundo ideal, todos los miembros de los jurados que utilizan una escala concreta coincidirían en sus juicios de cada caso. Cualquier desviación del juicio promediado cuenta como un error, y estos errores son la fuente del ruido del sistema.

Como también señalamos en el capítulo 6, el ruido del sistema puede dividirse en ruido de nivel y ruido de patrón. En este caso, el ruido de nivel es la variabilidad entre los miembros de los jurados en cuanto a la severidad en general. El ruido de patrón es la variabilidad, relativa a la media de un determinado miembro del jurado, en la forma en que ese miembro responde a casos particulares. Por tanto, podemos descomponer la varianza global de los juicios en tres elementos:

Varianza de los juicios = Varianza de las penas justas + (ruido de nivel)^2 + (ruido de patrón)^2

Este análisis, que descompone la varianza de las sentencias en tres términos, se llevó a cabo por separado para los tres juicios relativos a la indignación, la intención punitiva y la indemnización en dólares.

La figura 13 muestra los resultados.[6] La escala menos ruidosa es la de intención punitiva, en la que el ruido del sistema representa el 51 por ciento de la varianza; hay tanto ruido como justicia. La escala de indignación es claramente más ruidosa, 71 por ciento de ruido. Y la escala de dólares es, con mucho, la peor: ¡el 94 por ciento de la varianza en los juicios es ruido!

Las diferencias son sorprendentes porque las tres escalas son, en cuanto a su contenido, casi idénticas. Ya vimos que los valores justos de la indignación y la intención punitiva estaban casi completamente correlacionados, como insinúa la hipótesis de la indignación. Las valoraciones de la intención punitiva y las indemnizaciones en dólares responden exactamente a la misma pregunta —¿con qué severidad debe ser penada la empresa General Assistance?— en unidades diferentes. ¿Cómo podemos explicar las grandes diferencias que se muestran en la figura 13?

Probablemente estemos de acuerdo en que la de la indignación no es una escala muy precisa. Es cierto que existe un comportamiento «completamente aceptable», pero si hay un límite para enojarse con la General Assistance o con los otros demandados, ese límite es bastante vago. ¿Qué significa que un comportamiento sea «absolutamente indignante»? La falta de claridad en el extremo superior de la escala hace que sea inevitable algo de ruido.

La intención punitiva es más específica. «Pena severa» es más preciso que «absolutamente indignante», porque una «pena extremadamente severa» está limitada por el máximo prescrito por la ley. Puede que deseemos «castigar» al culpable, pero no podemos, por ejemplo, recomendar que se ejecute al director ejecutivo de General Assistance y a todo su equipo directivo. (Esperemos.) La escala de intención punitiva es menos ambigua porque su límite superior está más claramente especificado. Como era de esperar, también es menos ruidosa.

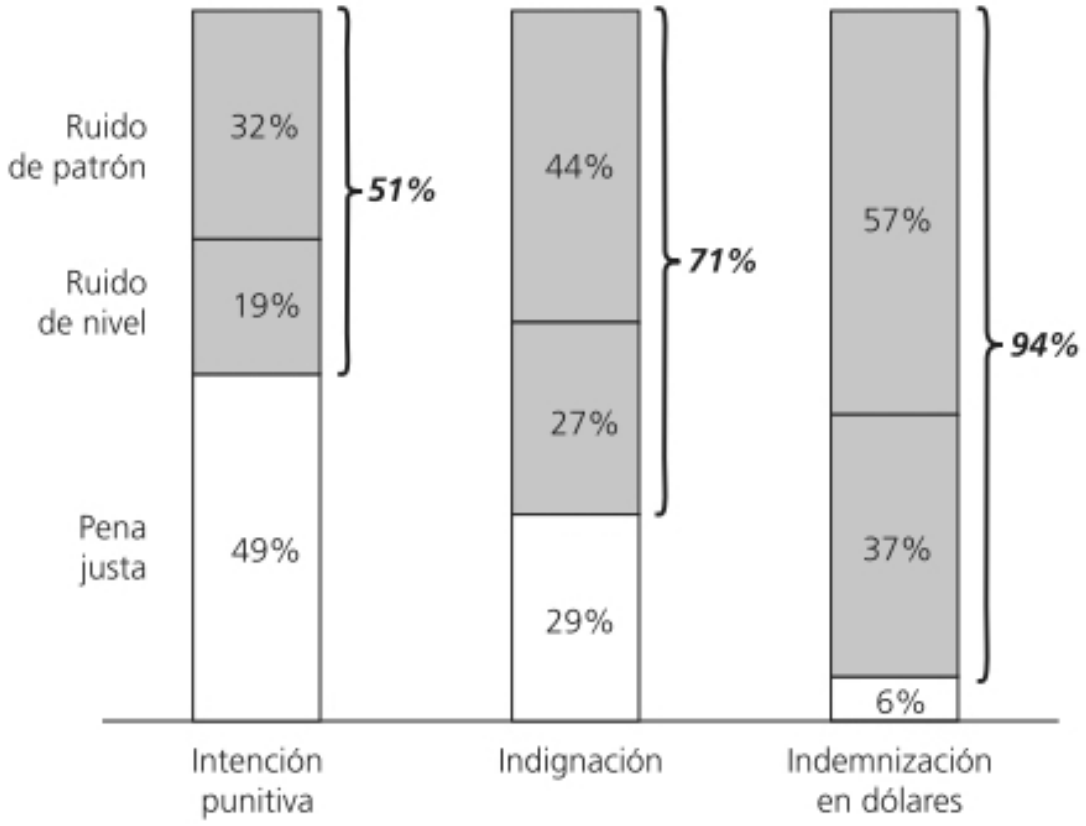

FIGURA 13. Componentes de la varianza en los juicios

La indignación y la intención punitiva se midieron en escalas de valoración similares, definidas más o menos claramente por etiquetas verbales. La escala de dólares pertenece a una familia diferente, que es mucho más problemática.

DÓLARES Y ANCLAJES

El título de nuestro trabajo académico expresa su mensaje central: «Shared Outrage and Erratic Awards. The Psychology of Punitive Damages». Hubo bastante coincidencia entre nuestros miembros de jurados experimentales en sus valoraciones de la intención punitiva; las valoraciones se explicaban en su mayoría por la indignación. Sin embargo, la cantidad de dólares era la que mejor simulaba la situación en los tribunales, y esta era inaceptablemente ruidosa.

La razón no es ningún misterio. Si el lector estableció una cantidad específica de dólares por los daños en el caso de Joan Glover, seguramente experimentó la sensación de que su elección de una cifra era esencialmente arbitraria. La sensación de arbitrariedad transmite una información importante: nos dice que otras personas tomarán decisiones arbitrarias muy diferentes y que los juicios serán muy ruidosos. Esto resulta ser una característica de la familia de escalas a la que pertenecen las indemnizaciones en dólares.

El legendario psicólogo de Harvard S. S. Stevens descubrió el sorprendente hecho de que las personas comparten poderosas intuiciones sobre las «ratios» de intensidad de muchas experiencias y actitudes subjetivas.[7] Pueden ajustar una luz para que parezca «el doble de brillante» que otra y están de acuerdo en que el significado emocional de una condena de diez meses de prisión no es ni de lejos diez veces peor que el de una condena de un mes. Stevens llamó a las escalas que se basan en esas intuiciones «escalas de ratio».

Se puede decir que nuestras intuiciones sobre el dinero se expresan en ratios por la facilidad con que entendemos expresiones como «¡Sara consiguió un aumento del 60 por ciento!» o «Nuestro vecino rico perdió la mitad de su fortuna de la noche a la mañana». La escala de dólares en las indemnizaciones punitivas es una escala de ratio para la medición de la intención de castigar. Como otras escalas de ratio, tiene un significativo cero (cero dólares) y no tiene un límite superior.

Stevens descubrió que una escala de ratio (como la escala de dólares) puede quedar amarrada a un único anclaje intermedio (el término de la jerga es *modulus*). En su laboratorio exponía a los observadores a una luz de cierta luminosidad con la instrucción de «dar a la luminosidad de esa luz un 10 (o 50, o 200) y asignar números a otras luminosidades partiendo de ese valor». Como se esperaba, los números que los observadores asignaron a las luces de diferente luminosidad fueron proporcionales al anclaje arbitrario que se les indicó. Un observador que estuviera anclado en el número 200 emitiría juicios con puntuaciones 20 veces más altas que si el anclaje hubiera estado en 10; la desviación estándar de los juicios del observador también sería proporcional al anclaje.

En el capítulo 13, describimos un divertido ejemplo de anclaje en el que la disposición de la gente a pagar por un objeto quedó muy determinada al

preguntarles primero si pagarían (en dólares) los dos últimos dígitos de su número de la Seguridad Social. Un resultado más sorprendente fue que el anclaje inicial también afectó a su disposición a pagar por toda una lista de otros objetos. Los participantes a los que se indujo a pagar una gran cantidad por un *trackball* inalámbrico aceptaron pagar una cantidad proporcionalmente mayor por un teclado inalámbrico. Parece que las personas son mucho más sensibles al valor «relativo» de bienes comparables que a su valor absoluto. Los autores del estudio llamaron al efecto persistente de un único anclaje «arbitrariedad coherente».[8]

Para apreciar el efecto de un anclaje arbitrario en el caso de Joan Glover, supongamos que el texto del principio de este capítulo incluyera la siguiente información:

> En un caso similar en que estuvo implicada otra compañía farmacéutica, la víctima, otra niña, sufrió un trauma psicológico moderado (como en la versión del daño leve antes referida). La indemnización punitiva se fijó en 1,5 millones de dólares.

Nótese que el problema de imponer una pena a General Assistance se ha vuelto de repente mucho más fácil. De hecho, es posible que ya se le haya ocurrido una cantidad al lector. Hay un multiplicador (o ratio) de las indemnizaciones en dólares que corresponde al contraste entre el grave daño que se infligió a Joan y el daño leve que sufrió la otra niña. Además, el único anclaje que ha leído (1,5 millones de dólares) es suficiente para anclar toda la escala de indemnizaciones en dólares. Ahora le resultará fácil fijar las indemnizaciones para casos más graves y más leves que los dos considerados hasta ahora.

Si se necesitan anclajes para emitir juicios en una escala de ratio, ¿qué ocurre cuando a los juzgadores no se les da un anclaje? Stevens halló la respuesta. En ausencia de una orientación proporcionada por el experimentador, las personas se ven obligadas a hacer una elección arbitraria cuando utilizan la escala por primera vez. A partir de ese momento, emiten sus juicios de forma coherente utilizando su primera respuesta como anclaje.

Es posible que el lector reconozca la tarea a la que se enfrentó al establecer las indemnizaciones en el caso de Joan Glover como un ejemplo de escala sin anclaje. Al igual que los observadores sin anclaje en el laboratorio de Stevens, tomó una decisión arbitraria sobre la pena correcta

para General Assistance. Los participantes en nuestro estudio sobre las indemnizaciones punitivas se enfrentaron al mismo problema: también se vieron obligados a tomar una decisión inicial arbitraria para el primer caso que vieron. Sin embargo, a diferencia del lector, no se limitaron a tomar una decisión arbitraria: siguieron estableciendo indemnizaciones punitivas para otros nueve casos. Estos nueve juicios no eran arbitrarios, porque podían ser coherentes con el juicio de anclaje inicial y, por tanto, entre sí.

Los descubrimientos del laboratorio de Stevens indican que el anclaje que producen los individuos debería tener un gran efecto en los valores absolutos de sus juicios posteriores referidos a los dólares, pero ningún efecto en las posiciones relativas de los diez casos. Un juicio inicial con una cuantía grande hace que en todos los demás juicios haya cuantías proporcionalmente grandes sin afectar a las cuantías relativas. Este razonamiento conduce a una conclusión sorprendente: aunque parezcan irremediablemente ruidosos, los juicios sobre los dólares reflejan, en realidad, las intenciones punitivas de los juzgadores. Para descubrir estas intenciones, solo tenemos que sustituir los valores absolutos en dólares por puntuaciones relativas.

Para contrastar esta idea, repetimos el análisis del ruido tras sustituir cada indemnización en dólares por su rango entre los diez juicios de un individuo. La indemnización más elevada en dólares se puntuó con 1, la siguiente más elevada con 2, y así sucesivamente. Esta transformación de las indemnizaciones en rangos elimina todos los errores de nivel de los miembros del jurado, ya que la distribución de rangos de 1 a 10 es la misma para cada uno, excepto para los empates ocasionales. (Por si el lector se lo pregunta, hemos de decir que había varias versiones del cuestionario, porque cada persona juzgaba diez de los veintiocho escenarios. Realizamos el análisis por separado para cada grupo de participantes que había respondido a los mismos diez escenarios, y presentamos un promedio.)

Los resultados fueron sorprendentes: la proporción de ruido en los juicios se redujo del 94 por ciento al 49 por ciento (figura 14). La transformación de las indemnizaciones en rangos reveló que los miembros de los jurados estaban de acuerdo en lo esencial con la pena apropiada en los diferentes casos.[9] De hecho, los rangos de las indemnizaciones fueron en todo caso algo menos ruidosos que las valoraciones originales de intención punitiva.

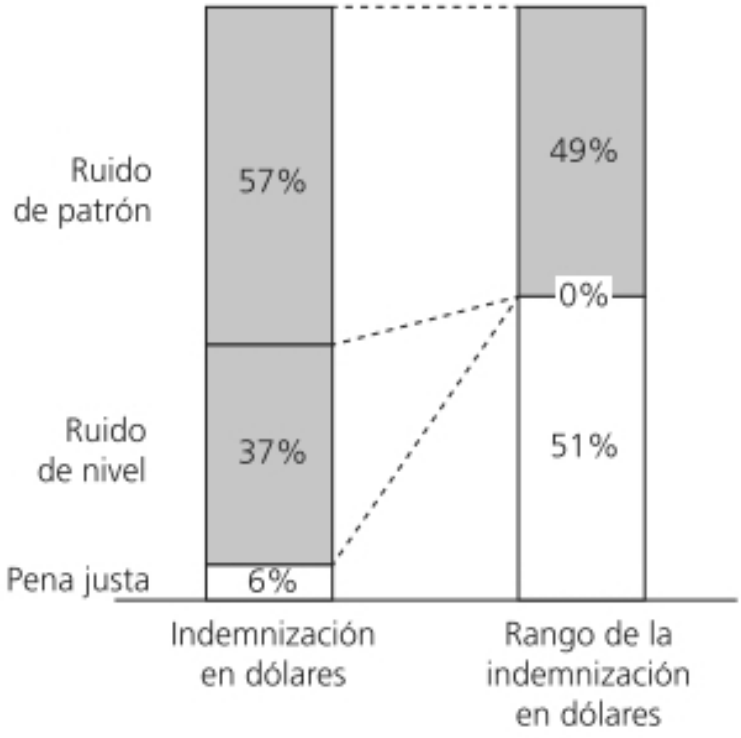

FIGURA 14. Ruido en el valor *versus* ruido en el rango

UNA CONCLUSIÓN DESAFORTUNADA

Los resultados son consistentes con la teoría que hemos esbozado: las indemnizaciones en dólares para todos los casos estaban ancladas en la cifra arbitraria que cada miembro del jurado eligió para el primer caso que vio. El rango relativo de los casos refleja las actitudes con bastante exactitud y, por tanto, no tiene mucho ruido, pero los valores absolutos de las indemnizaciones carecen esencialmente de sentido porque dependen de la cifra arbitraria elegida en el primer caso.

Paradójicamente, el caso que los miembros de los jurados evalúan en los juicios reales es el primero y único que ven. La práctica jurídica estadounidense requiere que los jurados civiles fijen una indemnización en dólares para un caso sin el beneficio de ningún anclaje que los guíe. La ley prohíbe de manera explícita cualquier comunicación al jurado de la cuantía de las indemnizaciones punitivas en otros casos. La suposición implícita en la ley es que el sentido de la justicia de los jurados les llevará directamente desde la consideración del delito hasta la pena adecuada. Esta suposición es

psicológicamente absurda, pues supone una capacidad que los humanos no poseen. Las instituciones de justicia deben reconocer las limitaciones de las personas que la administran.

El ejemplo de las indemnizaciones punitivas es extremo; los juicios profesionales rara vez se expresan en escalas tan completamente ambiguas. Sin embargo, las escalas ambiguas son comunes, lo que significa que el estudio de las indemnizaciones punitivas encierra dos lecciones generales aplicables en los negocios, la educación, los deportes, el Gobierno y otros ámbitos. En primer lugar, la elección de una escala puede suponer una gran diferencia en la cantidad de ruido de los juicios, porque las escalas ambiguas son ruidosas. En segundo lugar, es probable que la sustitución de los juicios absolutos por los relativos, cuando sea factible, reduzca el ruido.

SOBRE LAS ESCALAS

- «Hay mucho ruido en nuestros juicios. ¿Podría ser porque entendemos la escala de manera diferente?»
- «¿Podemos acordar un caso que sirva de anclaje y punto de referencia en la escala?»
- «Para reducir el ruido, ¿no deberíamos sustituir nuestros juicios por un rango?»

16
Patrones

¿Recuerda el lector a Julie, la niña precoz cuya calificación media en la universidad intentó adivinar en el capítulo 14? Le ofrecemos aquí una descripción más completa:

> Julie era hija única. Su padre era un abogado de éxito y su madre arquitecta. Cuando Julie tenía unos tres años, su padre contrajo una enfermedad autoinmune que lo obligó a trabajar en casa. Pasó mucho tiempo con Julie, y le enseñó pacientemente a leer. Ya leía con fluidez a los cuatro años. Su padre también intentó enseñarle aritmética, pero esa materia le resultaba difícil. Julie era una buena alumna en la escuela primaria, pero era emocionalmente dependiente y bastante impopular. Pasaba mucho tiempo sola, y llegó a ser una apasionada observadora de las aves gracias a un tío suyo, que le enseñó a distinguirlas.
>
> Sus padres se divorciaron cuando contaba once años, y eso se lo tomó muy mal. Sus notas empeoraron, y tenía frecuentes arrebatos en la escuela. En el instituto le fue muy bien en algunas asignaturas, como Biología y Escritura Creativa. Sorprendió a todos destacando en Física. Pero descuidó la mayoría de las otras asignaturas, y terminó sus estudios en el instituto con un notable.
>
> Al no ser admitida en las prestigiosas universidades en que quiso matricularse, Julie acabó en una buena escuela estatal, donde se especilizó en Estudios Medioambientales. Durante sus dos primeros años continuó con un patrón de frecuentes enredos emocionales, y fumaba hierba con bastante regularidad. Sin embargo, en su cuarto semestre, experimentó un intenso deseo de ingresar en la escuela de Medicina, y comenzó a tomarse sus estudios mucho más en serio.
>
> ¿Cuál es su mejor estimación de la calificación media de Julie?

PROBLEMAS: DIFÍCILES Y FÁCILES

Obviamente, este problema (llamémoslo Julie 2.0) se ha vuelto mucho más difícil. Todo lo que el lector conocía de Julie 1.0 era que sabía leer a los cuatro años. Con solo una pista, el poder del emparejamiento hizo el trabajo, y una estimación intuitiva de la nota media de Julie le vino rápidamente a la mente.

El emparejamiento seguiría operando como antes si el lector tuviera varios indicios que apuntaran en la misma dirección general. Por ejemplo, cuando este lee una descripción de Bill, el contable que toca jazz, toda su información («poco imaginativo», «bueno en matemáticas», «flojo en ciencias sociales») pintaba una imagen coherente y estereotipada.

Del mismo modo, si la mayor parte de los acontecimientos de la vida de Julie 2.0 fueran coherentes con una historia de precocidad y rendimientos superiores (con quizá unos pocos datos que indiquen un rendimiento simplemente «medio»), la tarea no le resultaría tan difícil. Cuando las pruebas disponibles dibujan una imagen coherente, nuestro pensamiento rápido del Sistema 1 no tiene ninguna dificultad para darle un sentido. Problemas de juicio tan sencillos como estos se resuelven fácilmente, y la mayoría de la gente está de acuerdo en su solución.

No es así con Julie 2.0. Lo que hace que este problema sea difícil es la presencia de múltiples señales contradictorias. Hay indicios de capacidad y de motivación, pero también de debilidades de carácter y rendimientos mediocres. La historia parece estar dispersa. No es fácil hallarle un sentido, porque los elementos no pueden encajar en una interpretación coherente. Por supuesto, la incoherencia no hace que la historia sea irreal o inverosímil. Nuestra vida es a menudo más compleja que las historias que nos gusta contar sobre ella.

Las señales múltiples y en contradicción crean la ambigüedad que define a los problemas difíciles de enjuiciar. La ambigüedad también explica por qué los problemas complejos son más ruidosos que los simples. La regla es sencilla: si hay más de una forma de ver algo, la gente variará su modo de verlo. La gente puede elegir diferentes pruebas para formar el núcleo de su relato, por lo que hay muchas conclusiones posibles. Si al lector le ha resultado difícil construir una historia que dé sentido a Julie 2.0, puede estar seguro de que otros lectores construirán historias diferentes que justifiquen juicios distintos del suyo. Esta es la variabilidad que produce el ruido de patrón.

¿Cuándo se siente uno confiado en un juicio? Deben cumplirse dos condiciones: la historia en la que cree debe ser enteramente coherente y no tiene que haber alternativas interesantes. La completa coherencia se consigue cuando todos los detalles de la interpretación elegida encajan en la historia y se refuerzan mutuamente. Por supuesto, también se puede lograr la coherencia, aunque de forma menos elegante, ignorando o justificando lo que no encaja. Lo mismo ocurre con las interpretaciones alternativas. El verdadero experto que ha «resuelto» un problema de juicio no solo sabe por qué su historia aclaratoria es correcta, sino que tiene la misma soltura para explicar por qué otras historias son erróneas. También en este caso una

persona puede obtener una confianza de igual fuerza, pero de peor calidad, si no considera alternativas o las suprime sin dilación.

La principal implicación de esta visión de la confianza es que la confianza subjetiva en el propio juicio no garantiza en absoluto el acierto. Además, la supresión de interpretaciones alternativas —un proceso bien documentado en la percepción—[1] podría inducir lo que hemos denominado la «ilusión de acuerdo» (véase el capítulo 2). Si las personas no pueden imaginar posibles alternativas a sus conclusiones, naturalmente supondrán que otros observadores también tienen que llegar a la misma conclusión. Por supuesto, pocos tenemos la suerte de estar muy seguros de todos nuestros juicios, y todos hemos tenido la experiencia de la incertidumbre, quizá tan reciente como la lectura sobre Julie 2.0. No todos tenemos mucha confianza en nosotros mismos en todo momento, pero la mayoría de las veces tenemos más confianza de la que deberíamos.[2]

RUIDO DE PATRÓN: ESTABLE O TRANSITORIO

Hemos definido un «error de patrón» como un error en el juicio de un individuo sobre un caso que no puede ser explicado por la suma de los efectos separados del caso y del juzgador. Un ejemplo extremo puede ser el de un juez normalmente indulgente que es excepcionalmente severo al condenar a un tipo concreto de acusado (por ejemplo, los que han cometido infracciones de tráfico). O, por ejemplo, el inversor normalmente cauto que abandona su habitual cautela cuando se le muestra el plan de una emocionante empresa emergente. Por supuesto, la mayoría de los errores de patrón no son extremos: observamos un error de patrón moderado en el juez indulgente que es menos indulgente de lo habitual cuando trata con reincidentes, o incluso más indulgente de lo habitual cuando condena a mujeres jóvenes.

Los errores de patrón nacen de una combinación de factores transitorios y permanentes. Los factores transitorios incluyen los que hemos descrito como «fuentes de ruido de ocasión», como el buen humor de un juez en el momento de la decisión o algún desafortunado suceso reciente que ocupe en ese momento la mente del juez. Otros factores son más permanentes; por ejemplo, el entusiasmo fuera de lo habitual de un empleador por las personas que asistieron a determinadas universidades o la inusitada

propensión de un médico a recomendar la hospitalización de personas con neumonía. Podemos escribir una ecuación sencilla que describe un error en un solo juicio:

Error de patrón = error de patrón estable + error transitorio (de ocasión)

Como el error de patrón estable y el error transitorio (de ocasión) son independientes y no están correlacionados, podemos ampliar la ecuación anterior para analizar sus varianzas:

$$(\text{Error de patrón})^2 = (\text{error de patrón estable})^2 + (\text{error de ocasión})^2$$

Como hicimos con otros componentes del error y del ruido, podemos representar gráficamente esta ecuación como una suma de cuadrados dispuestos en los lados de un triángulo rectángulo (figura 15).

Para un caso sencillo de ruido de patrón estable, consideremos a los contratadores que predicen el rendimiento futuro de ejecutivos sobre la base de un conjunto de calificaciones. En el capítulo 9 hablamos de un «modelo del juzgador». El modelo de un contratador individual asigna un peso a cada calificación, que corresponde a su importancia en los juicios de ese contratador. Las ponderaciones varían entre los contratadores: el liderazgo puede contar más para un contratador, y las habilidades de comunicación más para otro. Estas diferencias producen una variabilidad en la clasificación de los candidatos por parte de los contratadores (un ejemplo de lo que llamamos «ruido de patrón estable»).

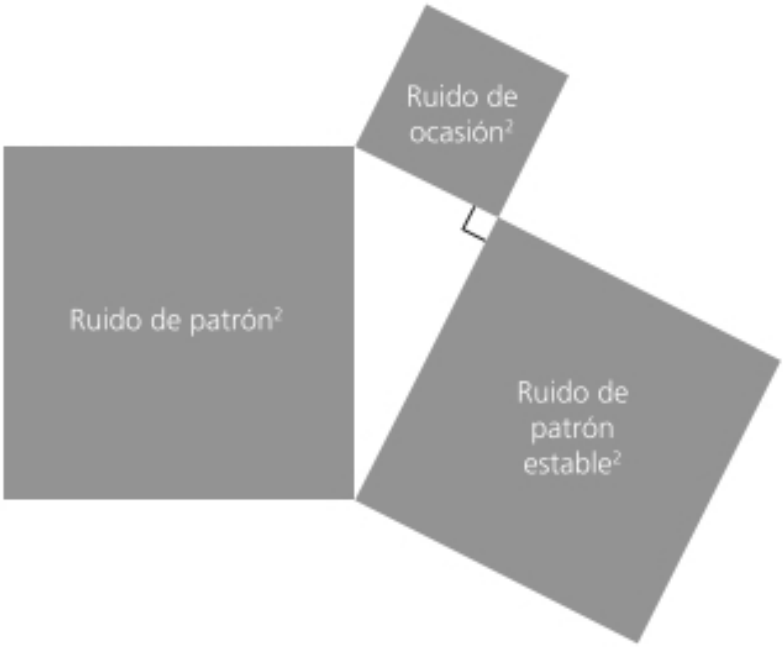

FIGURA 15. Descomposición del ruido de patrón

Las reacciones personales a casos individuales también pueden producir patrones que son estables, pero muy específicos. Considere el lector qué le llevó a prestar más atención a determinados aspectos de la historia de Julie que a otros. Algunos detalles del caso pueden resonar con una experiencia de su vida. Tal vez algo de Julie le recuerde a un pariente cercano que estuvo a punto de triunfar, pero que finalmente fracasó debido a lo que cree que son grandes defectos de carácter que fueron evidentes desde la adolescencia de ese pariente.

Por el contrario, la historia de Julie puede evocar el caso de una amiga cercana que, tras una adolescencia problemática, consiguió estudiar Medicina y ahora es una gran especialista. Las asociaciones que Julie evoca en diferentes personas son idiosincrásicas e impredecibles, pero es probable que sean estables: si el lector hubiera leído la historia de Julie la semana pasada, se habría acordado de las mismas personas y habría visto su historia bajo la misma y particular luz personal.

Las diferencias individuales en la calidad de los juicios son otra fuente de ruido de patrón. Imaginemos a una única analista con poderes de bola de cristal que nadie conoce (incluida ella misma). Su exactitud la haría

desviarse en muchos casos de la predicción media. En ausencia de datos de resultados, estas desviaciones se considerarían errores de patrón. Cuando los juicios no son verificables, una mayor precisión parecerá ruido de patrón.

El ruido de patrón nace también de diferencias sistemáticas en la capacidad de emitir juicios válidos sobre diferentes dimensiones de un caso. Considérese el proceso de selección de los equipos deportivos profesionales. Los entrenadores pueden centrarse en las habilidades en varios aspectos del juego, los médicos en la propensión a las lesiones y los psicólogos en la motivación y la resiliencia. Cuando estos diferentes especialistas evalúan a los mismos jugadores, podemos esperar una cantidad considerable de ruido de patrón. Del mismo modo, los profesionales que desempeñan la misma función generalista pueden ser más hábiles en algunos aspectos de la tarea de juicio que en otros.[3] En estos casos, el ruido de patrón se describe mejor como variabilidad en lo que las personas conocen que como error.

Cuando los profesionales toman decisiones por su cuenta, la variabilidad en las capacidades es simplemente ruido. Sin embargo, cuando los gestores tienen la oportunidad de formar equipos que lleguen a juicios de forma conjunta, la variedad de capacidades se convierte en un activo potencial, porque los diferentes profesionales cubrirán distintos aspectos del juicio y se complementarán entre sí.[4] En el capítulo 21 hablamos de esta oportunidad (y de lo que se necesita para aprovecharla).

En capítulos anteriores nos ocupamos de las dos loterías a las que se enfrenta el cliente de una compañía de seguros o el acusado al que se le asigna el juez que lo juzgará. Ahora podemos ver que la primera lotería, que elige a un profesional de un grupo de colegas, selecciona mucho más que el nivel medio de los juicios de ese profesional (el error de nivel). La lotería también selecciona un conjunto caleidoscópico de valores, preferencias, creencias, recuerdos, experiencias y asociaciones que son únicas de ese profesional en particular. Cada vez que alguien emite un juicio, lo hace también con su particular bagaje. Lo hace con los hábitos mentales que se formó en el ejercicio de la profesión y la sabiduría obtenida de los que fueron sus mentores. Se lleva consigo los éxitos que le hicieron ganar confianza y los errores que cuidadosamente procura no repetir. Y, en algún lugar de su cerebro, están las reglas formales que recuerda, las que olvidó y

las que aprendió que era mejor ignorar. Ninguna persona es exactamente igual a él en todos estos respectos; sus errores de patrón estables son exclusivos de él.

La segunda lotería es la que escoge el momento en que una persona emite su juicio, el estado de ánimo en que se encuentra y otras circunstancias ajenas que no deberían afectar a su juicio, pero que lo hacen. Esta lotería crea ruido de ocasión. Imagine el lector, por ejemplo, que poco antes de ver el caso de Julie, lee un artículo en el periódico sobre el consumo de drogas en los campus universitarios. El artículo contaba la historia de un estudiante bien dotado que estaba decidido a estudiar Derecho y se esforzaba mucho, pero no podía compensar el déficit que había acumulado por consumir drogas en sus primeros años de universidad. Como esta historia está fresca en su mente, prestará más atención al hábito de fumar marihuana de Julie en la evaluación de sus posibilidades generales. Sin embargo, probablemente no recordaría el artículo si se encontrara con la pregunta sobre Julie un par de semanas después (y obviamente no habría sabido nada de él si hubiera leído el caso ayer). El efecto de la lectura del artículo del periódico es transitorio; es ruido de ocasión.

Como ilustra este ejemplo, no hay una discontinuidad brusca entre el ruido de patrón estable y la variante inestable que llamamos «ruido de ocasión». La principal diferencia es si la sensibilidad única de una persona a algunos aspectos del caso es en sí misma permanente o transitoria. Cuando los desencadenantes del ruido de patrón están arraigados en nuestras experiencias y valores personales, podemos esperar que el patrón sea estable, un reflejo de nuestra singularidad.

LA ANALOGÍA DE LA PERSONALIDAD

La idea de la singularidad de las respuestas de determinadas personas a ciertas características o combinaciones de características no es intuitiva de inmediato. Para entenderlo, podríamos considerar otra compleja combinación de características que todos conocemos bien: la personalidad de la gente que nos rodea. El propio hecho de que un juzgador emita un juicio sobre un caso debe considerarse una cuestión particular de un tema más amplio como es el campo de la investigación de la personalidad: cómo

actúa una persona en una situación. Hay algo que aprender sobre el juicio tras las décadas de estudio intensivo del tema en general.

Los psicólogos llevan mucho tiempo intentando comprender y medir las diferencias individuales de personalidad. Las personas difieren unas de otras en muchos aspectos; el primer intento que se hizo de buscar en el diccionario los términos que pueden describir a una persona identificó dieciocho mil palabras.[5] En la actualidad, el modelo dominante de la personalidad, el modelo de los cinco grandes, combina rasgos en cinco agrupaciones (extraversión, amabilidad, responsabilidad, apertura a la experiencia, neuroticismo o inestabilidad emocional), y cada uno de los cinco grandes comprende un rango de rasgos distinguibles. Un rasgo de personalidad se entiende como un predictor de comportamientos reales. Si alguien se describe como escrupuloso, esperamos observar algunos comportamientos correspondientes (ser puntual, cumplir los compromisos, etcétera). Y si Pedro puntúa más alto que Juan en una medición de agresividad, observaríamos que, en la mayoría de las situaciones, Pedro se comporta más agresivamente que Juan. Sin embargo, la validez de los rasgos generales para predecir comportamientos específicos es bastante limitada; una correlación de .30 (PC = 60 %) se consideraría alta.[6]

El sentido común nos dice que, si bien el comportamiento puede derivarse de la personalidad, también se ve muy afectado por las «situaciones». En algunas de ellas nadie es agresivo y, en otras, todo el mundo lo es. Cuando consuelan a un amigo afligido, ni Pedro ni Juan actúan con agresividad; sin embargo, cuando ven un partido de fútbol, ambos mostrarán cierta agresividad. En definitiva —y esto no es nada sorprendente—, los comportamientos son una función de las personalidades *y* de las situaciones.

Lo que hace a las personas únicas y siempre interesantes es que esta unión de personalidad y situación no es una función mecánica y añadida. Por ejemplo, las situaciones que despiertan más o menos agresividad no son las mismas para todas las personas. Aunque Pedro y Juan sean igual de agresivos por término medio, no muestran forzosamente la misma agresividad en todos los contextos. Quizá Pedro sea agresivo con sus compañeros, pero dócil con sus superiores, mientras que el nivel de agresividad de Juan no es sensible al nivel jerárquico. Tal vez Juan sea muy

propenso a la agresividad cuando se le critica, pero tienda a contenerse cuando se le amenaza físicamente.[7]

Es probable que estos patrones característicos de respuesta a las situaciones sean bastante estables a lo largo del tiempo. Constituyen gran parte de lo que consideramos la «personalidad» de alguien, aunque no se prestan a una descripción por un rasgo amplio. Pedro y Juan pueden compartir la misma puntuación en un test de agresividad, pero son únicos en su patrón de respuesta a los desencadenantes y los contextos de agresividad. Dos personas que comparten un nivel de un rasgo —si, por ejemplo, son igual de obstinadas o igual de generosas— deben ser descritas por dos distribuciones de comportamientos que tienen la misma media, pero no necesariamente el mismo patrón de respuestas a distintas situaciones.

Ahora puede verse el paralelismo entre esta discusión sobre la personalidad y el modelo de juicio que hemos presentado. Las diferencias de nivel entre los juzgadores se corresponden con las diferencias entre las puntuaciones sobre rasgos de la personalidad, las cuales representan una media de los comportamientos en múltiples situaciones. Los casos son análogos a las situaciones. El juicio de una persona sobre un problema concreto solo es moderadamente predecible a partir del nivel medio de esa persona, del mismo modo que los comportamientos específicos solo son moderadamente predecibles a partir de rasgos de la personalidad. La clasificación de los individuos según su juicio varía considerablemente de un caso a otro, porque las personas difieren en su reacción a las características y combinaciones de características que encuentran en cada caso. Lo que distingue a un individuo que emite juicios y toma decisiones es un patrón único de sensibilidad a las características y un patrón también único en el juicio de los casos.

La singularidad personal es normalmente un motivo de celebración, pero este libro se ocupa de los juicios profesionales, donde la variación es problemática y el ruido es un error. El sentido de la analogía es que el ruido de patrón en los juicios no es aleatorio, aunque tengamos pocas esperanzas de explicarlo y aunque los individuos que emiten juicios característicos no puedan explicarlos.

SOBRE EL RUIDO DE PATRÓN

- «Pareces estar seguro de tu conclusión, pero este no es un problema fácil: hay indicios que apuntan en diferentes direcciones. ¿Has pasado por alto interpretaciones alternativas de las pruebas?»
- «Tú y yo hemos entrevistado al mismo candidato, y normalmente somos entrevistadores igual de exigentes. Sin embargo, tenemos juicios muy diferentes. ¿De dónde viene este ruido de patrón?»
- «La singularidad personal es lo que hace a las personas capaces de innovar y ser creativas, así como interesantes y emocionantes. Sin embargo, cuando se trata de juzgar, esa singularidad no es una ventaja.»

17
Las fuentes de ruido

Esperamos que, a estas alturas, el lector esté de acuerdo en que, dondequiera que haya juicio, hay ruido. También esperamos que, en su caso, no haya más de lo que cree. Este mantra sobre el ruido nos motivó cuando empezamos nuestro proyecto, pero nuestra forma de pensar sobre el asunto ha evolucionado a lo largo de años de trabajo. A continuación, repasaremos las principales lecciones que hemos aprendido sobre los componentes del ruido, sobre su importancia respectiva en el cuadro general del ruido y sobre el lugar que ocupa el ruido en el estudio del juicio.

LOS COMPONENTES DEL RUIDO

La figura 16 ofrece una representación gráfica combinada de las tres ecuaciones que hemos introducido en los capítulos 5, 6 y 16. La figura ilustra tres desgloses sucesivos del error:

- Error en sesgo y ruido del sistema.
- Ruido del sistema en ruido de nivel y ruido de patrón.
- Ruido de patrón en ruido de patrón estable y ruido de ocasión.

Ahora podemos ver cómo el ECM se descompone en los cuadrados del sesgo y de los tres componentes del ruido de los que hemos hablado.[1]

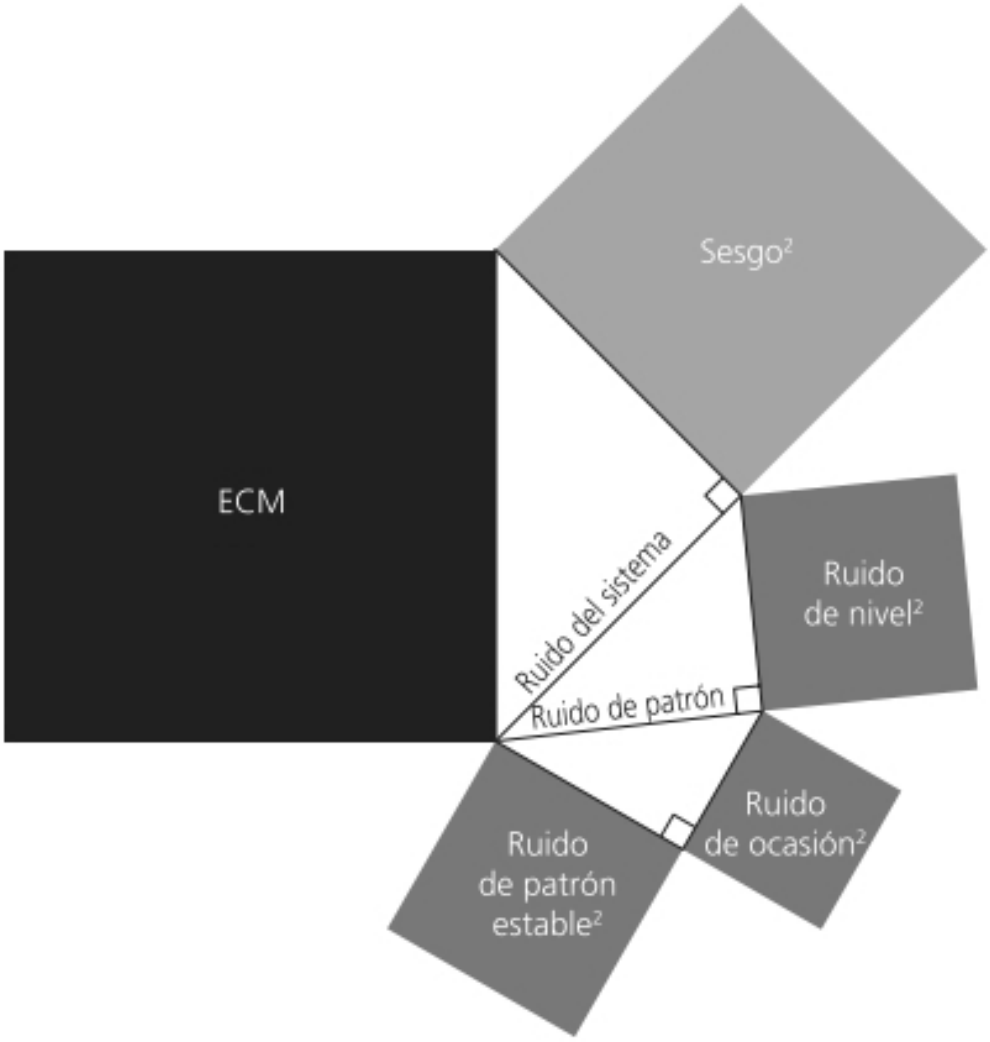

FIGURA 16. Error, sesgo y los componentes del ruido

Cuando iniciamos nuestra investigación, nos centramos en los pesos relativos del sesgo y el ruido en el error total. Pronto llegamos a la conclusión de que el ruido es a menudo un componente de error mayor que el sesgo, y sin duda merecía la pena estudiarlo con más detalle.

Nuestras primeras reflexiones sobre los componentes del ruido se guiaron por la estructura de complejas auditorías del ruido en las que varias personas emiten juicios individuales sobre múltiples casos. El estudio de los jueces federales fue un ejemplo y otro el de las indemnizaciones punitivas. Los datos de estos estudios nos proporcionaron estimaciones sólidas del ruido de nivel. Por otra parte, como cada participante juzga cada caso, pero

solo lo hace una vez, no hay manera de saber si el error residual, que hemos llamado «error de patrón», es transitorio o estable. En el espíritu conservador del análisis estadístico, el error residual se denomina comúnmente «término de error» y se trata como aleatorio. En otras palabras, la interpretación por defecto del ruido de patrón es que consiste enteramente en ruido de ocasión.

Esta interpretación convencional del ruido de patrón como error aleatorio limitó nuestro pensamiento durante mucho tiempo. Parecía natural centrarse en el ruido de nivel, es decir, en las continuas diferencias entre los jueces severos y los indulgentes, o entre los analistas optimistas y los pesimistas. También nos intrigaban las pruebas de la influencia en los juicios de las circunstancias irrelevantes y transitorias que crean ruido de ocasión.

Las pruebas nos llevaron de forma gradual a darnos cuenta de que los juicios ruidosos que emiten las diferentes personas están determinados en gran medida por algo que no es un sesgo general ni un sesgo transitorio y aleatorio del individuo: las persistentes reacciones personales de individuos particulares a una gran cantidad de características que determinan sus reacciones a casos específicos. Finalmente, llegamos a la conclusión de que debíamos abandonar nuestra suposición por defecto sobre la naturaleza transitoria del ruido de patrón.

Aunque queremos tener cuidado de no generalizar en exceso a partir de lo que sigue siendo una selección limitada de ejemplos, los estudios que hemos reunido, tomados en conjunto, sugieren que el ruido de patrón estable es realmente más significativo que los demás componentes del ruido del sistema. Dado que rara vez disponemos de un cuadro completo de los componentes del error en un mismo estudio, se requiere cierta triangulación para formular esta conclusión provisional. En resumen, esto es lo que sabemos (y lo que no).

CALIBRACIÓN DE LOS COMPONENTES

En primer lugar, tenemos varias estimaciones de los pesos relativos del ruido de nivel y del ruido de patrón. En general, parece que el ruido de patrón contribuye más que el ruido de nivel. En la compañía de seguros del capítulo 2, por ejemplo, las diferencias entre los agentes de seguros en la media de las primas que fijaban solo representaban el 20 por ciento del

ruido total del sistema; el 80 por ciento restante era ruido de patrón. Entre los jueces federales del capítulo 6, el ruido de nivel (diferencias en severidad media) representaba algo menos de la mitad del ruido total del sistema; el ruido de patrón era el mayor componente. En el experimento de las indemnizaciones punitivas, la cantidad total de ruido del sistema variaba ampliamente en función de la escala utilizada (intención punitiva, indignación o indemnizaciones en dólares), pero la parte de ruido de patrón en ese total era más o menos constante: representó el 63 por ciento, 62 por ciento y 61 por ciento del ruido total del sistema para las tres escalas utilizadas en el estudio. Otros estudios que revisaremos en la quinta parte del libro, en concreto sobre las decisiones de contratación de personal, son coherentes con esta conclusión provisional.

El hecho de que, en estos estudios, el ruido de nivel no sea, por lo general, el mayor componente del ruido del sistema es ya un mensaje importante, porque el ruido de nivel es la única forma de ruido que las organizaciones pueden controlar (a veces) sin llevar a cabo auditorías del ruido. Cuando los casos se asignan de forma más o menos aleatoria a profesionales individuales, las diferencias en el nivel medio de sus decisiones demuestran el ruido de nivel. Por ejemplo, los estudios sobre las oficinas de patentes observaron grandes diferencias en la propensión media de los examinadores a conceder patentes, con los consiguientes efectos en la incidencia de los litigios sobre estas patentes.[2] Del mismo modo, los responsables de los casos en los servicios de protección de la infancia varían en su tendencia a enviar a los niños a centros de acogida, con consecuencias a largo plazo para el bienestar de los niños.[3] Estas observaciones se basan únicamente en una estimación del ruido de nivel. Si hay más ruido de patrón que ruido de nivel, estos resultados, ya de por sí llamativos, subestiman la magnitud del problema del ruido por lo menos en un factor de dos. (Hay excepciones a esta regla provisional. La escandalosa variabilidad de las decisiones de los jueces de asilo político se debe, casi con toda seguridad, al ruido de nivel más que al ruido de patrón, que sospechamos que también es grande.[4])

El siguiente paso es analizar el ruido de patrón separando sus dos componentes. Hay buenas razones para suponer que el ruido de patrón estable, y no el ruido de ocasión, es el componente dominante. La auditoría

de las sentencias de los jueces federales ilustra nuestro razonamiento. Comencemos con la posibilidad extrema de que todo el ruido de patrón sea transitorio. En ese supuesto, las sentencias serían inestables e inconsistentes a lo largo del tiempo hasta un punto que nos parecería inverosímil: tendríamos que esperar que la diferencia media entre los juicios «del mismo caso por el mismo juez» en diferentes ocasiones fuera de unos 2,8 años.[5] La variabilidad de la media de las sentencias entre los jueces es ya chocante. La misma variabilidad en las sentencias de un juez individual en las distintas ocasiones sería absurda. Parece más razonable concluir que los jueces difieren en sus reacciones ante distintos acusados y diferentes delitos, y que estas diferencias son muy personales, pero estables.

Para cuantificar con mayor precisión qué parte del ruido de patrón es estable y qué parte es ruido ocasional, necesitamos estudios en los que los mismos jueces realicen dos evaluaciones independientes de cada caso. Como hemos señalado, la obtención de dos juicios independientes es en general imposible en los estudios del juicio, porque es difícil garantizar que el segundo juicio de un caso sea realmente independiente del primero. Sobre todo cuando el juicio es complejo, hay una alta probabilidad de que el individuo reconozca el problema y repita el juicio original.

Un grupo de investigadores de Princeton dirigido por Alexander Todorov ha diseñado ingeniosas técnicas experimentales para superar este problema.[6] Reclutaron a participantes de Amazon Mechanical Turk, un sitio en el que los individuos prestan servicios a corto plazo, como responder a cuestionarios, y se les paga por su tiempo. En un experimento, los participantes vieron imágenes de rostros (generados por un programa informático, pero completamente indistinguibles de los rostros de las personas reales) y los calificaron en función de varios atributos, como la simpatía y la honradez. El experimento se repitió, con las mismas caras y los mismos encuestados, una semana después.

Es razonable esperar que haya menos consenso en este experimento que en los juicios profesionales, como los de los jueces que dictan sentencias. Todo el mundo puede estar de acuerdo en que algunas personas son muy atractivas y otras nada atractivas, pero, en un rango significativo, esperamos que las reacciones a los rostros sean en gran medida idiosincrásicas. De hecho, hubo poco acuerdo entre los observadores: en las estimaciones de la honradez, por ejemplo, las diferencias entre las imágenes representaron solo

el 18 por ciento de la varianza de los juicios. El 82 por ciento restante de la varianza era ruido.

También es razonable esperar una menor estabilidad en estos juicios, ya que la calidad de los juicios realizados por los participantes a los que se les paga por responder a las preguntas en línea suele ser considerablemente menor que en los entornos profesionales. Sin embargo, el mayor componente del ruido fue el ruido de patrón estable. El segundo componente más importante del ruido fue el ruido de nivel, es decir, las diferencias entre los observadores en sus estimaciones medias de la honradez. El ruido de ocasión, aunque seguía siendo importante, era el componente menor.

Los investigadores llegaron a las mismas conclusiones cuando pidieron a los participantes que hicieran otros juicios (sobre preferencias entre coches o alimentos, por ejemplo, o sobre cuestiones más cercanas a lo que llamamos «juicios profesionales»). Por ejemplo, en una repetición del estudio sobre indemnizaciones punitivas que analizamos en el capítulo 15, los participantes valoraron su intención punitiva en diez casos de daños personales en dos ocasiones distintas separadas por una semana. También en este caso, el ruido de patrón estable fue el componente más importante. En todos estos estudios, los individuos no coincidieron entre sí, pero se mantuvieron bastante estables en sus juicios. Esta «consistencia sin consenso», en palabras de los investigadores, proporciona una clara prueba del ruido de patrón estable.

La prueba más contundente del papel de los patrones estables procede del amplio estudio sobre jueces y libertad bajo fianza que mencionamos en el capítulo 10.[7] En una parte de este excepcional estudio, los autores crearon un modelo estadístico que simulaba cómo cada juez utilizaba los indicios disponibles para decidir si concedía la libertad bajo fianza. Crearon modelos a medida de ciento setenta y tres jueces. A continuación, emplearon los jueces simulados para que tomaran decisiones sobre 141.833 casos, con lo que obtuvieron ciento setenta y tres decisiones para cada caso: un total de más de veinticuatro millones de decisiones.[8] A petición nuestra, los autores llevaron a cabo generosamente un análisis especial en el que separaron los juicios de varianza en tres componentes: la varianza «verdadera» de las decisiones medias para cada uno de los casos, el ruido

de nivel creado por las diferencias entre los jueces en su propensión a conceder la libertad bajo fianza y el ruido de patrón restante.

Este análisis es relevante para nuestra hipótesis, porque el ruido de patrón, tal y como se mide en este estudio, es totalmente estable. La variabilidad aleatoria del ruido de ocasión no está representada, pues se trata de un análisis de «modelos» que predicen la decisión de un juez. Solo se incluyen las reglas de predicción individuales que son estables y verificables.

La conclusión fue clara: este ruido de patrón estable era casi cuatro veces mayor que el ruido de nivel (el ruido de patrón estable representaba el 26 por ciento de la varianza total y el ruido de nivel el 7 por ciento de esta).[9] Los patrones de juicio individuales, estables e idiosincrásicos que pudieron identificarse fueron en general mucho mayores que las diferencias de severidad.

Todas estas pruebas concuerdan con la investigación sobre el ruido ocasional que revisamos en el capítulo 7: mientras que la existencia del ruido ocasional es sorprendente y hasta inquietante, nada indica que la variabilidad en la misma persona sea mayor que las diferencias entre personas. El componente más importante del ruido del sistema es el que teníamos descuidado al principio: el ruido de patrón estable, la variabilidad entre jueces en sus juicios de casos particulares.

Dada la relativa escasez de investigaciones relevantes, nuestras conclusiones son provisionales, pero reflejan un cambio en la forma de entender el ruido (y el modo de abordarlo). Al menos en principio, el ruido de nivel —las simples diferencias entre los magistrados globalmente consideradas— debería ser un problema relativamente fácil de medir y de abordar. Si hay calificadores anormalmente «duros», funcionarios de custodia infantil «cautos», o agentes de crédito con «aversión al riesgo», las organizaciones que los emplean podrían tratar de igualar el nivel medio de sus juicios. Las universidades, por ejemplo, abordan este problema cuando exigen a los profesores que se atengan a una distribución predeterminada de las calificaciones dentro de cada clase.

Por desgracia, como ahora nos damos cuenta, centrarse en el ruido de nivel supone perder gran parte de lo que son las diferencias individuales. El ruido no es en su mayor parte producto de diferencias de nivel, sino de interacciones: cómo tratan los distintos jueces a determinados acusados,

cómo tratan los distintos profesores a determinados alumnos, cómo tratan los distintos trabajadores sociales a determinadas familias, cómo tratan los distintos líderes determinadas visiones del futuro. El ruido es sobre todo un subproducto de nuestra singularidad, de nuestra «personalidad juzgadora». Reducir el ruido de nivel sigue siendo un objetivo que merece la pena, pero alcanzar solo este dejaría sin solución a la mayor parte del problema que supone el ruido del sistema.

EXPLICANDO EL ERROR

Hemos visto que hay mucho que decir sobre el ruido, pero el tema está casi ausente del todo de la conciencia pública y de los debates sobre el juicio y el error. A pesar de lo evidente de su presencia y de los múltiples mecanismos que lo producen, el ruido rara vez se menciona como un factor importante en los juicios. ¿Cómo puede suceder esto? ¿Por qué nunca invocamos el ruido para explicar los malos juicios, mientras que habitualmente culpamos a los sesgos? ¿Por qué es tan excepcional pensar en el ruido como fuente de error, a pesar de su omnipresencia?

La clave de este misterio es que, si bien la media de los errores (el sesgo) y la variabilidad de los errores (el ruido) desempeñan papeles equivalentes en la ecuación de error, pensamos en ellos de maneras radicalmente diferentes. Y nuestra forma habitual de dar sentido al mundo que nos rodea hace casi imposible reconocer el papel del ruido.

Anteriormente, hemos señalado que, en retrospectiva, podemos entender con facilidad los acontecimientos, aunque no podamos predecirlos antes de que sucedan. En el valle de lo normal, estos no nos sorprenden y les hallamos fácil explicación.

Lo mismo puede decirse de los juicios. Al igual que otros acontecimientos, los juicios y las decisiones suelen darse en el valle de lo normal; generalmente no nos sorprenden. Por un lado, los juicios que producen resultados satisfactorios son normales y rara vez se cuestionan. Cuando el lanzador elegido para el tiro libre marca puntos, cuando la operación de corazón tiene éxito o cuando una empresa de nueva creación prospera, suponemos que las razones que tuvieron los que tomaron la decisión fueron las correctas. Al fin y al cabo, se demostró que tenían

razón. Como cualquier otra historia poco sorprendente, una historia de éxito se explica por sí misma una vez que se conoce el resultado.

En cambio, sentimos la necesidad de explicar los resultados anormales: los malos y, en ocasiones, los sorprendentemente buenos (como la pésima apuesta empresarial que da buenos resultados). Las explicaciones que apelan al error o a un don especial son mucho más populares de lo que merecen, porque las apuestas importantes del pasado se convierten fácilmente en actos de genialidad o de locura cuando se conoce su resultado. Un sesgo psicológico bien documentado, denominado «error fundamental de atribución», es una fuerte tendencia a atribuir la culpa, o el mérito, a los agentes por acciones y resultados que se explican mejor por la suerte o por circunstancias objetivas. Otro sesgo, la retrospectiva, distorsiona los juicios de modo que resultados que no podían preverse parecen muy previsibles si echamos la vista atrás.

Las explicaciones para errores de juicio no son difíciles de hallar; encontrar las razones de los juicios es, en todo caso, más fácil que hallar las causas de los acontecimientos. Siempre podemos invocar las razones de las personas que emiten los juicios. Si eso no es suficiente, podemos culpar a su incompetencia. Y, en las últimas décadas, se ha generalizado otra explicación de los malos juicios: el sesgo psicológico.

Un abundante cuerpo de investigación en psicología y economía de la conducta ha documentado una larga lista de sesgos psicológicos: la falacia de la planificación, el exceso de confianza, la aversión a la pérdida, el efecto de dotación, el sesgo del *statu quo*, el excesivo descuento del futuro («sesgo del presente») y muchos otros, incluidos, por supuesto, los sesgos a favor o en contra de diversas categorías de personas. Mucho se sabe sobre las condiciones en las que cada uno de estos sesgos puede influir en los juicios y las decisiones, y se conocen bastantes métodos que permitirían a un observador de la toma de decisiones reconocer el pensamiento sesgado en tiempo real.

Un sesgo psicológico es una explicación causal legítima de un error de juicio en el caso de que el sesgo se hubiera predicho de antemano o se hubiera detectado en tiempo real. Un sesgo psicológico que se identifique solo después del hecho puede seguir proporcionando una explicación útil, aunque sea provisional, si también ofrece una predicción sobre el futuro. Por ejemplo, el sorprendente rechazo de una firme candidata a un puesto

puede sugerir una hipótesis más general de sesgo relativo al sexo que los futuros nombramientos del mismo comité confirmarán o refutarán. Considérese, por el contrario, una explicación causal que solo se aplique a un acontecimiento: «En aquel caso fallaron, por lo que debieron de tener un exceso de confianza». La afirmación carece por completo de contenido, pero proporciona una ilusión de comprensión que puede ser bastante satisfactoria. El profesor de la escuela de negocios Phil Rosenzweig ha argumentado de forma convincente que las explicaciones huecas en términos de sesgos son habituales en los debates sobre los resultados empresariales.[10] Su popularidad atestigua la necesidad imperante de historias causales que den sentido a la experiencia.

EL RUIDO ES ESTADÍSTICO

Como señalamos en el capítulo 12, nuestra forma normal de pensar es causal. Atendemos naturalmente a lo particular, siguiendo y creando historias causalmente coherentes sobre casos individuales en los que los fallos suelen atribuirse a errores y los errores a sesgos. La facilidad con que se pueden explicar los malos juicios no deja espacio para el ruido en nuestras consideraciones sobre los errores.

La invisibilidad del ruido es una consecuencia directa del pensamiento causal. El ruido es inherentemente estadístico: solo se hace visible cuando pensamos estadísticamente en un conjunto de juicios similares. De hecho, es difícil no verlo: es la variabilidad en las estadísticas retrospectivas sobre decisiones de condenas y primas de seguros; es el rango de posibilidades cuando consideramos cómo predecir un resultado futuro; es la dispersión de los impactos en el blanco. Causalmente, el ruido no está en ninguna parte; estadísticamente, está en todas.

Por desgracia, adoptar la perspectiva estadística no es fácil. Invocamos sin esfuerzo las causas de los acontecimientos que observamos, pero pensar estadísticamente en ellas es algo que debe aprenderse y siempre requerirá un esfuerzo. Las causas son naturales; la estadística es difícil.

El resultado es un marcado desequilibrio en la forma de ver el sesgo y el ruido como fuentes de error. Quien haya asistido a algún curso de introducción a la psicología, seguramente recordará las ilustraciones en las que una figura destacada y rica en detalles sobresale de un fondo indistinto.

Nuestra atención está firmemente fijada en la figura, incluso cuando es pequeña contra el fondo. Las demostraciones de figura-fondo son una metáfora adecuada para nuestras intuiciones sobre el sesgo y el ruido: el sesgo es una figura convincente, mientras que el ruido es el fondo al que no prestamos atención. Así es como seguimos sin ser conscientes de un gran defecto en nuestros juicios.

SOBRE LAS FUENTES DE RUIDO

- «Vemos fácilmente las diferencias en el nivel medio de los juicios, pero ¿cuál es la magnitud del ruido de patrón que no vemos?»
- «Dices que la causa de este juicio fueron los sesgos, pero ¿dirías lo mismo si el resultado hubiera sido diferente? ¿Y puedes saber si hubo ruido?»
- «Nos centramos, con razón, en la reducción de los sesgos. Preocupémonos también de reducir el ruido.»

QUINTA PARTE

Mejorar los juicios

¿Cómo puede una organización mejorar los juicios que hacen sus profesionales? Y, en particular, ¿cómo puede una organización reducir el ruido de los juicios? ¿Cómo respondería un juzgador a estas preguntas?

Un primer paso necesario es conseguir que la organización reconozca que el ruido en los juicios profesionales es un problema que merece atención. Para llegar a este punto, recomendamos una auditoría del ruido (véase el apéndice A para una descripción detallada). En una auditoría del ruido, varios individuos juzgan los mismos problemas. El ruido es la variabilidad de esos juicios. Habrá casos en los que esta variabilidad pueda atribuirse a la incompetencia: algunos juzgadores saben de qué hablan; otros, no. Cuando existe este tipo de deficiencias (ya sea en general o en determinados tipos de casos), la prioridad debe ser, por supuesto, mejorar las capacidades deficientes. Sin embargo, como hemos visto, puede haber una gran cantidad de ruido incluso en los juicios de profesionales competentes y bien formados.

Si merece la pena encarar la cantidad de ruido del sistema, sustituir el juicio por reglas o algoritmos es una opción que debería considerarse, ya que eliminará el ruido por completo. Pero las reglas tienen sus propios problemas (como veremos en la sexta parte), y hasta los defensores más entusiastas de la inteligencia artificial coinciden en que los algoritmos no son, ni serán pronto, un sustituto universal del juicio humano. La tarea de mejorar el juicio urge más que nunca y es el tema de esta parte del libro.

Una forma sensata de mejorar los juicios es, por supuesto, seleccionar los mejores juzgadores humanos posibles. En el campo de tiro, algunos tiradores tienen una puntería muy buena. Igual ocurre con cualquier tarea de juicio profesional: los más capacitados serán menos ruidosos y acusarán menos sesgos. La manera de buscar los mejores juzgadores es en ocasiones obvia: si alguien quiere resolver un problema de ajedrez, debe preguntar a un gran maestro, no a los autores de este libro. Sin embargo, en la mayoría de los problemas, las características de los juzgadores superiores son más difíciles de discernir. Estas son el tema del capítulo 18.

A continuación, analizaremos los métodos para la reducción de los errores en los juicios. Los sesgos psicológicos están implicados tanto en el sesgo estadístico como en el ruido. Como veremos en el capítulo 19, ha habido muchos intentos de contrarrestar los sesgos psicológicos, con algunos fracasos claros y algunos éxitos evidentes. Repasaremos brevemente las estrategias de eliminación de sesgos y propondremos un método prometedor que, hasta donde sabemos, no ha sido analizado de forma sistemática: pedir a un «observador de decisiones» previamente designado que busque signos diagnósticos que puedan indicar en tiempo real que el trabajo de un grupo está siendo afectado por uno o varios sesgos conocidos. El apéndice B ofrece un ejemplo de lista de control de sesgos que podría utilizar un observador de decisiones.

A continuación, pasaremos a nuestro principal objetivo en esta parte del libro: la lucha contra el ruido. Introduciremos el tema de la «higiene de las decisiones», el método que recomendamos para reducir el ruido en los juicios humanos. Presentaremos estudios de casos en cinco campos diferentes. En cada campo, examinaremos la frecuencia del ruido y algunas de las historias de terror que genera. También revisaremos el éxito —o la falta de éxito— de los intentos de reducir el ruido. En cada campo se utilizarán, por supuesto, múltiples enfoques, pero, con el fin de facilitar la exposición, cada capítulo hará hincapié en una única estrategia de higiene de las decisiones.

En el capítulo 20 comenzaremos con el caso de la ciencia forense, que ilustra la importancia de «secuenciar información». La búsqueda de coherencia lleva a las personas a formarse unas primeras impresiones basadas en las escasas pruebas disponibles y, posteriormente, a confirmar su prejuicio cuando este aparece. Por eso es importante no exponerse a información irrelevante al principio del proceso de juzgar.

En el capítulo 21 abordaremos el caso de la predicción, que ilustra el valor de una de las estrategias más importantes de reducción del ruido: la «suma de múltiples juicios independientes». El principio de la «sabiduría de la multitud» se basa en el promedio de múltiples juicios independientes, lo que garantiza la reducción del ruido. Más allá de la media directa, existen otros métodos para sumar juicios, también ilustrados con el ejemplo de la predicción.

El capítulo 22 ofrecerá una revisión del ruido en la medicina y de los intentos de reducirlo. Señalará la importancia y la aplicabilidad general de una estrategia de reducción del ruido que ya introdujimos con el ejemplo de las sentencias penales: las «directrices de juicio». Las directrices pueden ser un poderoso mecanismo de reducción de ruido, porque reducen directamente la variabilidad entre jueces en las sentencias finales.

En el capítulo 23, nos ocuparemos de un reto familiar en la vida empresarial: las evaluaciones del rendimiento. Los intentos de reducir el ruido demuestran la crucial importancia de utilizar una «escala compartida basada en una visión externa». Se trata de una importante estrategia de higiene de las decisiones por una sencilla razón: el juicio implica la traslación de una impresión a una escala, y, si distintos jueces utilizan escalas diferentes, habrá ruido.

El capítulo 24 examinará el tema, relacionado pero distinto, de la selección de personal, que ha sido ampliamente investigado en los últimos cien años y que ilustra el valor de una estrategia esencial para la higiene de las decisiones: la «estructuración de juicios complejos». Por «estructuración» entendemos la descomposición de un juicio en sus componentes, la gestión del proceso de recogida de datos para garantizar que los *inputs* sean independientes entre sí y la dilación del debate holístico y el juicio final hasta que no se hayan recogido todos esos *inputs*.

Nos basamos en las lecciones aprendidas en el campo de la selección de personal para proponer, en el capítulo 25, un enfoque general de la evaluación de opciones denominado «protocolo de evaluaciones mediadoras» (o PEM para abreviar). El PEM parte de la premisa de que «las opciones son como los candidatos», y describe de forma esquemática cómo la toma de decisiones estructurada, junto con las demás estrategias de higiene de las decisiones arriba mencionadas, pueden introducirse en un proceso de decisión típico tanto para las decisiones recurrentes como para las singulares.

Una cuestión general antes de embarcarnos: sería valioso poder especificar, incluso cuantificar, los beneficios probables de cada estrategia de higiene de las decisiones en distintos contextos. También sería valioso saber cuál de las estrategias es más beneficiosa y cómo compararlas. Cuando se controla el flujo de información, ¿en qué medida se reduce el ruido? Si el objetivo es reducir el ruido en la práctica, ¿cuántos juicios

deben sumarse? La estructuración de juicios puede ser valiosa, pero ¿cuál es su valor exacto en diferentes contextos?

Dado que el tema del ruido ha atraído poca atención, estas siguen siendo preguntas abiertas que, con el tiempo, la investigación podría abordar. A efectos prácticos, las ventajas de una u otra estrategia dependerán del entorno concreto en el que se utilicen. Considérese la adopción de directrices: a veces producen enormes beneficios (como veremos en algunos diagnósticos médicos). Sin embargo, en otros entornos, los beneficios de la adopción de directrices podrían ser modestos, quizá porque no haya mucho ruido para empezar o quizá porque incluso las mejores directrices posibles no reducen mucho el error. En cualquier contexto, quien toma decisiones debe aspirar a lograr una comprensión más precisa de los probables beneficios de cada estrategia de higiene de las decisiones (y de los costes correspondientes, que analizaremos en la sexta parte).

18
Mejores juzgadores para mejores juicios

Hasta ahora hemos hablado principalmente de los juzgadores humanos sin distinguir entre ellos. Sin embargo, es obvio que en cualquier tarea que requiera juicio, algunas personas lo harán mejor que otras. Incluso una suma de juicios basada en la sabiduría de la multitud puede ser mejor si la multitud está compuesta por personas más capaces.[1] Una cuestión importante es, por tanto, cómo identificar a esos jueces mejores.

Hay tres cuestiones importantes. Los juicios son menos ruidosos y menos sesgados cuando, quienes los emiten, están bien formados, son más inteligentes y tienen el estilo cognitivo adecuado. En otras palabras: los buenos juicios dependen de lo que se sabe, de lo bien que se piensa y de «cómo» se piensa. Los buenos juzgadores suelen ser experimentados e inteligentes, pero también suelen tener la mente activamente abierta y estar dispuestos a aprender de nuevas informaciones.

EXPERTOS Y EXPERTOS DE RESPETO

Es casi tautológico decir que la capacidad de los juzgadores afecta a la calidad de sus juicios. Los radiólogos expertos, por ejemplo, tienen más probabilidades de diagnosticar correctamente una neumonía, y en la predicción de aconteceres mundiales hay «superanalistas» que superan con seguridad a otros analistas no tan superiores. Si se reúne un grupo de abogados que son verdaderos especialistas en algún área del derecho, es probable que hagan predicciones similares, y buenas, sobre el resultado de los litigios comunes en los tribunales. Las personas altamente cualificadas son menos ruidosas, y también acusan menos sesgos.

Estas personas son verdaderos expertos en las tareas en cuestión. Su superioridad sobre otros se puede comprobar gracias a la disponibilidad de los datos de resultados. Al menos en principio, podemos elegir un médico, un analista o un abogado en función de la frecuencia con la que haya acertado en el pasado. (Por razones obvias, esto puede ser difícil en la

práctica; no recomendamos que intente someter a su médico de cabecera a un examen de competencia.)

Como también hemos señalado, muchos juicios no son comprobables. Dentro de ciertos límites, no podemos conocer fácilmente ni definir sin ninguna duda el valor real al que los juicios apuntarían. La evaluación de seguros y las sentencias penales entran en esta categoría, al igual que la cata de vinos, la calificación de redacciones, la crítica de libros y de películas y otros innumerables juicios. Sin embargo, hay profesionales de estos ámbitos a los que llamamos «expertos». La confianza que tenemos en el juicio de estos últimos se basa por entero en el respeto de que gozan entre sus colegas. Los llamaremos «expertos de respeto».

La expresión «experto de respeto» no pretende ser irrespetuosa. El hecho de que algunos expertos no se sometan a una evaluación del acierto de sus juicios no es una crítica; es una realidad en muchos campos. Muchos profesores, académicos y consultores son expertos de respeto. Su credibilidad depende del respeto de sus alumnos, compañeros o clientes. En todos estos campos, y en muchos más, los juicios de un profesional solo pueden compararse con los de sus colegas.

A falta de verdaderos valores que determinen quién tiene o no razón, solemos valorar la opinión de los expertos de respeto incluso cuando no están de acuerdo. Imagínese, por ejemplo, un elenco de analistas políticos con perspectivas muy diferentes sobre lo que ha provocado una crisis diplomática y cómo se desarrollará. (Este desacuerdo no es infrecuente; no sería un elenco muy interesante si todos estuvieran de acuerdo.) Todos los analistas creen que hay un planteamiento correcto y que el suyo es el que más se acerca a él. Cuando los escuchamos, podemos encontrar a varios de los analistas igualmente impresionantes, y a sus argumentos igualmente convincentes. No podemos saber cuál de ellos tiene razón (y puede que ni siquiera sepamos después si no habrían formulado sus análisis como predicciones claramente verificables). Sabemos que al menos algunos de los analistas estarán equivocados porque no hay acuerdo entre ellos. Sin embargo, los respetamos como expertos.

O consideremos un conjunto diferente de expertos que no hacen predicciones. Tres filósofos morales, todos ellos bien formados, están reunidos en una sala. Uno de ellos sigue a Immanuel Kant; otro, a Jeremy Bentham, y un tercero, a Aristóteles. Con todo el respeto que reclama la

moral, discrepan apasionadamente. La cuestión puede ser si es legítimo mentir y cuándo, o los derechos de los animales, o la finalidad de las leyes penales. Los escuchamos con atención. Podemos admirar la claridad y precisión de su pensamiento. Y tendemos a estar de acuerdo con un filósofo, pero los respetamos a todos.

¿Por qué? De forma más general, ¿cómo personas respetadas por la calidad de su juicio deciden confiar en alguien como experto, cuando no hay datos para establecer objetivamente su aptitud? ¿Qué es lo que hace de alguien un experto de respeto?

Parte de la respuesta es la existencia de normas compartidas, o de doctrina profesional. Los expertos suelen obtener cualificaciones profesionales de comunidades profesionales, y reciben formación y supervisión en sus organizaciones. Los médicos que completan su residencia y los jóvenes abogados que aprenden de un socio veterano no solo aprenden a usar las herramientas técnicas de su oficio; están formados para utilizar ciertos métodos y seguir ciertas normas.

Las normas compartidas dan a los profesionales una idea de los elementos que deben tener en cuenta y de cómo emitir y justificar sus juicios finales. En la compañía de seguros, por ejemplo, los tasadores de reclamaciones no tenían ninguna dificultad para ponerse de acuerdo y describir las consideraciones pertinentes que debían incluirse en una lista de control para evaluar una reclamación.

Por supuesto, este acuerdo no impidió que los tasadores de reclamaciones variasen ampliamente en sus evaluaciones, porque la doctrina no especifica con claridad cómo proceder. No hay una receta que pueda aplicarse mecánicamente. En cambio, la doctrina deja espacio a la interpretación. Los expertos siguen haciendo juicios, no cálculos. De ahí que el ruido sea inevitable. Incluso profesionales con idéntica formación que están de acuerdo con la doctrina que aplican se distanciarán unos de otros en su aplicación.

Más allá del conocimiento de las normas compartidas, también es necesaria la experiencia. Alguien puede ser un joven prodigio si su especialidad es el ajedrez, los conciertos de piano o el lanzamiento de jabalina, porque los resultados validan su nivel de competencia. Sin embargo, los agentes de seguros, los dactiloscopistas o los jueces suelen

necesitar algunos años de experiencia para tener credibilidad. No hay jóvenes prodigios entre los agentes de seguros.

Otra característica de los expertos de respeto es su capacidad para emitir y explicar sus juicios seguros de sí mismos. Tendemos a confiar más en las personas que confían en sí mismas que en las que muestran sus dudas. La heurística de la confianza apunta al hecho de que, en un grupo, las personas seguras de sí mismas tienen más peso que otras, aunque no tengan motivos para estar tan seguras.[2] Los expertos de respeto destacan en la construcción de historias coherentes. Su experiencia les permite reconocer patrones, razonar por analogía con casos anteriores y formular y confirmar hipótesis con rapidez. Acomodan fácilmente los hechos que conocen en una historia coherente que inspira confianza.

INTELIGENCIA

La formación, la experiencia y la seguridad permiten a los expertos de respeto inspirar confianza. Sin embargo, estos atributos no garantizan la calidad de sus juicios. ¿Cómo podemos saber qué expertos pueden emitir buenos juicios?

Hay sobradas razones para creer que la inteligencia general puede estar asociada a un mejor juicio. La inteligencia está correlacionada con un buen rendimiento en casi todos los ámbitos. En igualdad de condiciones, se asocia no solo a un mayor éxito académico, sino también a un mayor desempeño laboral.[3]

En las discusiones sobre la medición de la inteligencia o de la capacidad intelectual general (CIG, término que ahora se suele preferir al de cociente intelectual o CI), abundan los debates y los malentendidos. Persisten las concepciones erróneas sobre la naturaleza innata de la inteligencia, pero, de hecho, los test miden las capacidades desarrolladas, que son en parte una función de los rasgos hereditarios y en parte están influidas por el ambiente, incluidas en este las oportunidades educativas.[4] A muchas personas les preocupan los efectos adversos de la selección basada en la CIG en grupos sociales identificables y la legitimidad de utilizar los test de CIG con fines de selección.

Tenemos que separar estas preocupaciones sobre el uso de los test de la realidad de su valor predictivo. Desde que el ejército estadounidense

empezó a utilizar pruebas de capacidad intelectual hace más de un siglo, miles de estudios han medido la relación entre las puntuaciones de los test cognitivos y el rendimiento posterior. El mensaje que se desprende de este cúmulo de investigaciones es inequívoco. Como dice una reseña, «la CIG predice tanto el nivel ocupacional alcanzado como el rendimiento dentro de la ocupación elegida, y lo hace mejor que cualquier otra capacidad, rasgo o disposición y mejor que la experiencia laboral».[5] Naturalmente, otras capacidades cognitivas son asimismo importantes (más adelante hablaremos de ello). También lo son muchos rasgos de la personalidad, como la diligencia y las «agallas» (*grit*), definidas como la perseverancia y la pasión en la persecución de objetivos a largo plazo.[6] Y sí, hay varias formas de inteligencia que los test de CIG no miden, como la inteligencia práctica y la creatividad. Los psicólogos y los neurocientíficos distinguen entre la inteligencia cristalizada, la capacidad de resolver problemas basándose en un almacén de conocimientos sobre el mundo (incluidas las operaciones aritméticas) y la inteligencia fluida, la capacidad de resolver problemas novedosos.[7]

Sin embargo, a pesar de su tosquedad y sus limitaciones, la CIG, medida utilizando test estandarizados que contienen preguntas sobre problemas verbales, cuantitativos y espaciales, sigue siendo, con mucho, el único mejor predictor de resultados importantes. Como añade la revisión antes mencionada, el poder predictivo de la CIG es «superior que la mayoría de lo hallado en la investigación psicológica».[8] La solidez de la asociación entre la capacidad intelectual general y el éxito laboral aumenta, lógicamente, con la complejidad del trabajo en cuestión: la inteligencia importa más para los científicos de cohetes que para los que cumplen tareas más sencillas. En los trabajos de alta complejidad, las correlaciones que se observan entre las puntuaciones de las pruebas estandarizadas y el rendimiento se sitúan en el rango de .50 (PC = 67 %).[9] Como hemos señalado, una correlación de .50 indica un valor predictivo muy grande conforme a los estándares de las ciencias sociales.[10]

Sobre todo en las discusiones sobre juicios profesionales cualificados, una objeción importante y frecuente a la relevancia de las mediciones de la inteligencia es que probablemente todos los que hacen tales juicios sean individuos de elevada CIG. Los médicos, los jueces o los agentes de

seguros más veteranos están mucho más educados que la población en general, y es muy probable que obtengan una puntuación mucho más alta en cualquier medición de la capacidad cognitiva. Se podría creer razonablemente que una CIG elevada no supone una gran diferencia entre ellos, que es simplemente el billete de entrada al grupo de personas con elevado éxito, no la fuente de las diferencias de éxito dentro de ese grupo.

Esta creencia, aunque extendida, es incorrecta. No cabe duda de que el rango de CIG que se encuentra en una determinada ocupación es más amplio en la parte inferior del rango de ocupaciones que en la superior: hay individuos con una CIG alta en ocupaciones de nivel inferior, pero casi no hay individuos con una CIG inferior a la media entre abogados, químicos o ingenieros.[11] Desde esta perspectiva, la capacidad intelectual alta es por lo visto una condición necesaria para acceder a las profesiones de elevado estatus.

Sin embargo, esta medida no capta las diferencias de éxito «dentro» de estos grupos. Incluso entre el 1 por ciento de las personas con mayor capacidad cognitiva (evaluada a los trece años), los resultados excepcionales están fuertemente correlacionados con la CIG. En comparación con los que están en el cuartil inferior de este 1 por ciento superior, los que están en el cuartil superior tienen entre dos y tres veces más probabilidades de obtener un título de doctorado, publicar un libro u obtener una patente. En otras palabras, no solo importa la diferencia de CIG entre el percentil 99 y el 80 o el 50, sino que sigue importando —¡y mucho! — entre el percentil 99,88 y el 99,13.[12]

Otra llamativa ilustración del vínculo entre la capacidad y los resultados es la que aporta un estudio de 2013 centrado en los directores ejecutivos (CEO) de las empresas de la lista Fortune 500 y los cuatrocientos veinticuatro multimillonarios estadounidenses (el 0,0001 por ciento de los estadounidenses ricos). Como era predecible, se encontró que estos grupos hiperelitistas estaban formados por personas pertenecientes al colectivo de las personas más capacitadas intelectualmente. Sin embargo, el estudio también encontró que, «dentro» de estos grupos, los niveles de educación y capacidad más elevados estaban relacionados con una mayor remuneración (en el caso de los CEO) y un mayor patrimonio neto (en el caso de los multimillonarios). Por cierto, que los famosos que abandonan la universidad y se convierten en multimillonarios, como Steve Jobs, Bill Gates y Mark

Zuckerberg, son los árboles que ocultan el bosque: considerando que alrededor de un tercio de los adultos estadounidenses han obtenido un título universitario, el 88 por ciento de los multimillonarios lo poseen.[13]

La conclusión es clara. La CIG contribuye de forma significativa a la calidad del rendimiento en las ocupaciones que requieren juicio, incluso dentro de un grupo de individuos de alta capacidad. La idea de que existe un umbral a partir del cual la CIG deja de marcar una diferencia no está respaldada por ninguna prueba. Esta conclusión sugiere a su vez que, si los juicios profesionales no son verificables, pero se supone que apuntan a una diana invisible, es más probable que los juicios de las personas de alta capacidad estén cerca de ella. Si hay que elegir a personas para que emitan juicios, escoger a las que tienen la mayor capacidad intelectual tiene mucho sentido.

Sin embargo, esta línea de razonamiento tiene una importante limitación. Como no es posible hacer pruebas estandarizadas a todo el mundo, hay que adivinar quiénes son las personas con mayor CIG. Y una CIG alta mejora el rendimiento en muchos frentes, incluida la capacidad de convencer a los demás de que se tiene razón. Los individuos con alta capacidad intelectual tienen más probabilidades que otros de emitir mejores juicios y ser verdaderos expertos, pero también poseen más probabilidades de impresionar a sus compañeros, ganarse la confianza de los demás y convertirse en expertos de respeto en ausencia de cualquier retroalimentación de la realidad. Los astrólogos medievales debieron de estar entre los individuos de mayor CIG de su época.

Puede ser sensato depositar nuestra confianza en personas que parecen inteligentes y que pueden articular de manera convincente la lógica de sus juicios, pero esta estrategia es insuficiente, y hasta puede ser contraproducente. ¿Existen, entonces, otras formas de identificar a verdaderos expertos? ¿Tienen las personas que emiten los mejores juicios otros rasgos reconocibles?

ESTILO COGNITIVO

Independientemente de la capacidad intelectual, las personas difieren en su «estilo cognitivo», o en su planteamiento de las tareas de juicio. Se han desarrollado muchos instrumentos para captar los estilos cognitivos. La

mayoría de sus mediciones se correlacionan con la CIG (y entre sí), pero miden cosas diferentes.

Una de estas mediciones es el «test de reflexión cognitiva» (TRC), que se hizo famoso por la ya omnipresente pregunta sobre la pelota y el bate: «Un bate y una pelota cuestan 1,10 dólares en total. El bate cuesta 1,00 dólar más que la pelota. ¿Cuánto cuesta la pelota?». Entre otras preguntas que se han propuesto para medir la reflexión cognitiva se encuentra esta: «Si estás corriendo una carrera y pasas a la persona que va en segundo lugar, ¿en qué lugar estás?».[14] Las preguntas del TRC intentan medir la probabilidad de que la gente descarte la primera respuesta (errónea) que se le ocurra («diez céntimos» para la pregunta de la pelota y el bate, y «primero» para el ejemplo de la carrera). Las puntuaciones más bajas del TRC se asocian a muchos juicios y creencias que encontramos en el mundo real, como la creencia en los fantasmas, la astrología y la percepción extrasensorial.[15] Las puntuaciones predicen si la gente se creerá las *fake news* descaradamente falsas.[16] Incluso se asocian a la frecuencia con que la gente usará sus *smartphones*.[17]

Muchos consideran que el TRC es un instrumento para medir un concepto más amplio: la propensión a utilizar procesos de pensamiento reflexivos frente a los impulsivos.[18] Sencillamente ocurre que algunas personas prefieren reflexionar detenidamente, mientras que otras, ante el mismo problema, tienden a confiar en sus primeros impulsos. En nuestra terminología, el TRC puede verse como una medida de la propensión de las personas a confiar en el pensamiento lento (del Sistema 2), más que en el pensamiento rápido (del Sistema 1).

Se han desarrollado otras autoevaluaciones para medir esta propensión (y todas estas pruebas se hallan, por supuesto, intercorrelacionadas). La escala de necesidad de conocimiento, por ejemplo, pregunta a las personas si tienden a pensar mucho en los problemas.[19] Para obtener una puntuación alta en la escala, tendrían que estar de acuerdo con esta frase: «Tiendo a marcarme objetivos que solo pueden lograrse realizando un esfuerzo mental considerable» y no estarlo con esta otra: «No me paro mucho a pensar». Las personas con una gran necesidad de conocimiento tienden a ser menos susceptibles a sesgos cognitivos conocidos.[20] También se han registrado algunas asociaciones más extrañas: quien evita que le cuenten el argumento y el final de las películas es probable que tenga una gran necesidad de

conocimiento, y quien tiene una escasa necesidad de conocimiento prefiere que le cuenten el argumento y el final.[21]

Dado que esta escala es una autoevaluación y que la respuesta socialmente deseable es bastante obvia, la escala plantea cuestiones pertinentes. Es poco probable que a alguien que trate de impresionar se le ocurra decir: «No me paro mucho a pensar». Por esa razón, otros test tratan de medir habilidades en lugar de utilizar las autodescripciones.

Un ejemplo es la escala de competencia adulta en la toma de decisiones, que mide la propensión de las personas a cometer errores de juicio típicos, como el exceso de confianza o el carácter contradictorio en la percepción del riesgo.[22] Otra es la evaluación del pensamiento crítico de Halpern, que se centra en la capacidad para el pensamiento crítico e incluye tanto una disposición al pensamiento racional como un conjunto de habilidades que pueden aprenderse. Para esta evaluación, se hacen preguntas como la siguiente: «Imagine que un amigo le pide consejo sobre cuál de dos dietas de adelgazamiento debe elegir. Una informa de que los clientes pierden una media de 11 kilos, y la otra informa de que pierden una media de 14 kilos. ¿Qué preguntas le gustaría que le respondiera antes de elegir una de las dos dietas?». Quien respondiese, por ejemplo, que desearía saber cuántas personas han perdido ese peso y si han conservado esa pérdida de peso durante un año o más sumaría puntos por aplicar el pensamiento crítico. Las personas que obtienen una buena puntuación en la escala de competencia adulta en la toma de decisiones, o en la evaluación de Halpern, parecen hacer mejores juicios en su vida: sufren menos acontecimientos vitales adversos provocados por malas elecciones, como tener que pagar con retraso el alquiler de una película o sufrir un embarazo no deseado.[23]

Sería razonable suponer que todas estas mediciones de estilo y aptitud cognitiva —y muchas otras— predicen generalmente los juicios. Sin embargo, su relevancia parece variar en función de la tarea. Cuando Uriel Haran, Ilana Ritov y Barbara Mellers buscaron estilos cognitivos que pudieran predecir la capacidad de predicción, descubrieron que la necesidad de conocimiento no predecía quién se esforzaría más en buscar nueva información. Tampoco encontraron que la necesidad de conocimiento se asociara de manera fiable con un mejor rendimiento.[24]

La única medición de estilo cognitivo o personalidad que hallaron para predecir la capacidad para hacer pronósticos fue otra escala, desarrollada

por el profesor de Psicología Jonathan Baron, para medir el «pensamiento de mentalidad activamente abierta».[25] Tener una mente abierta es buscar activamente información que contradiga las propias hipótesis preexistentes. Esa información incluye las opiniones discrepantes de otros y la cuidadosa ponderación de nuevas pruebas frente a antiguas creencias. Las personas de mente abierta están de acuerdo con afirmaciones como esta: «Dejarse convencer por un argumento contrario es un signo de buen carácter». Y no están de acuerdo con la afirmación de que «cambiar de opinión es un signo de debilidad», o la de que «la intuición es la mejor guía para tomar decisiones».

En otras palabras: mientras que las puntuaciones de reflexión cognitiva y necesidad de conocimiento miden la propensión a un pensamiento lento y cuidadoso, el pensamiento abierto activo va más allá. Es la humildad de los que son siempre conscientes de que su juicio es un trabajo incesante y anhelan ser corregidos. En el capítulo 21 veremos que este estilo de pensamiento caracteriza a los mejores analistas, que cambian continuamente de opinión y revisan sus creencias en respuesta a nueva información. Curiosamente, hay algunas pruebas de que el pensamiento abierto activo es una capacidad que se puede enseñar.[26]

No pretendemos sacar aquí conclusiones definitivas sobre cómo elegir a los individuos que emitirán buenos juicios en un ámbito determinado. Sin embargo, de este breve repaso se desprenden dos principios generales. En primer lugar, es prudente reconocer la diferencia entre los ámbitos en los que la experiencia puede confirmarse mediante la comparación con los valores reales (como la previsión meteorológica) y los ámbitos que son competencia de los expertos de respeto. Un analista político puede parecer elocuente y convincente, y un gran maestro de ajedrez puede parecer tímido e incapaz de explicar el razonamiento que hay detrás de algunas de sus jugadas. Sin embargo, probablemente deberíamos tomar el juicio profesional del primero con más escepticismo que el del segundo.

En segundo lugar, algunos jueces son mejores que sus colegas igual de cualificados y experimentados. Si son mejores, es menos probable que estén sesgados o sean ruidosos. Entre las muchas cosas que explican estas diferencias, la inteligencia y el estilo cognitivo son importantes. Aunque ninguna medición o escala predice de forma inequívoca la calidad de los juicios, podemos tratar de distinguir al tipo de personas que buscan

activamente nueva información que pueda contradecir sus creencias previas, que son metódicas a la hora de integrar esa información en su perspectiva actual y que, como consecuencia, están dispuestas, incluso ansiosas, de cambiar de opinión.

La personalidad de las personas de juicio excelente puede no encajar en el estereotipo generalmente aceptado de un líder decidido. La gente suele confiar y querer a los líderes que son firmes y claros y que parecen saber, de forma inmediata y con inquebrantable certeza, lo que conviene. Estos líderes inspiran confianza. Sin embargo, las pruebas nos enseñan que, si el objetivo es reducir el error, es mejor que los líderes (y otros) estén abiertos a los contraargumentos y sepan que pueden estar equivocados. Si acaban mostrándose firmes, será al final de un proceso, no al principio.

SOBRE LOS MEJORES JUZGADORES

- «Usted es un experto. Pero ¿sus juicios son verificables, o es usted un experto de respeto?»
- «Tenemos que elegir entre dos opiniones, y no sabemos nada de la experiencia y la trayectoria de estas personas. Sigamos el consejo de la más inteligente.»
- «Sin embargo, la inteligencia es solo una parte de la historia. También es importante “cómo” se piensa. Tal vez deberíamos elegir a la persona más reflexiva y de mente más abierta, en lugar de a la más inteligente.»

19
Supresión de sesgos e higiene de las decisiones

Muchos investigadores y organizaciones han perseguido el objetivo de suprimir los sesgos en los juicios. Este capítulo examina sus principales conclusiones. Distinguiremos entre diferentes tipos de intervenciones destinadas a esa supresión y analizaremos una que merece una ulterior investigación. A continuación, pasaremos a la reducción del ruido e introduciremos la idea de «higiene de las decisiones».[1]

INTERVENCIÓN *EX POST* Y *EX ANTE*

Una buena manera de caracterizar los dos métodos principales de supresión de sesgos es volver a la analogía de la medición. Supongamos que alguien sabe que la báscula de su baño añade, de promedio, 150 gramos a su peso. La escala está sesgada. Sin embargo, esto no la hace inútil. Puede asumir el sesgo de una de las dos maneras posibles: puede corregir cada lectura de su poco amable báscula restando 150 gramos. Sin duda, esto puede resultar un poco pesado (y puede que se le olvide hacerlo). Una alternativa sería ajustar el indicador y mejorar la precisión del instrumento de una vez por todas.

Estos dos métodos de corrección de mediciones sesgadas tienen análogos directos en las intervenciones para suprimir sesgos en los juicios. Actúan *ex post*, corrigiendo los juicios después de que se hayan emitido, o *ex ante*, interviniendo antes de hacer un juicio o tomar una decisión.

La intervención *ex post*, o correctiva, suele realizarse de forma intuitiva. Supongamos que alguien está supervisando un equipo encargado de un proyecto, y que el equipo calcula que puede completar su proyecto en tres meses. Tal vez quiera añadir un amortiguador al juicio de los miembros del equipo y planificar para cuatro meses o más, corrigiendo así un sesgo (la falacia de la planificación) que supone está presente.

Este tipo de corrección de sesgos se realiza a veces de forma más sistemática. En el Reino Unido, el Ministerio de Hacienda ha publicado un «libro verde», una guía para evaluar programas y proyectos.[2] El libro

insta a los planificadores a hacer frente a los sesgos optimistas aplicando ajustes porcentuales a las estimaciones del coste y duración de un proyecto. Lo ideal es que estos ajustes se basen en los niveles históricos de sesgo optimista de una organización. Si no se dispone de estos datos históricos, el «libro verde» recomienda aplicar porcentajes de ajuste genéricos para cada tipo de proyecto.

Las intervenciones *ex ante*, o preventivas, se clasifican a su vez en dos grandes categorías. Algunas de las más prometedoras están diseñadas para modificar el entorno en el que tiene lugar el juicio o la decisión. Estas modificaciones, o *nudges*, como se conocen, tratan de reducir el efecto de los sesgos o hacer listas de sesgos para tomar una mejor decisión.[3] Un ejemplo sencillo es la inscripción automática en planes de pensiones. Diseñada para superar la inercia, la postergación y el sesgo optimista, la inscripción automática garantiza que los empleados ahorrarán para su jubilación a menos que opten de forma deliberada por no hacerlo. La inscripción automática ha demostrado ser sumamente eficaz para aumentar las tasas de participación. A veces, el programa va acompañado de planes como el de Ahorra Más Mañana, por los que los empleados pueden acordar destinar un determinado porcentaje de sus futuros aumentos salariales al ahorro. La inscripción automática puede utilizarse en muchos terrenos (por ejemplo, la inscripción automática en la energía verde, en los programas de comidas escolares gratuitas para niños pobres o en otros planes de prestaciones).

Otros *nudges* actúan sobre diferentes aspectos de la arquitectura de la elección. Pueden hacer que la decisión correcta sea la más fácil (por ejemplo, reduciendo las cargas administrativas para tener acceso a determinados cuidados por problemas de salud mental). O pueden destacar ciertas características de un producto o una actividad (por ejemplo, haciendo explícitas y claras unas tasas antes ocultas). Las tiendas de comestibles y los sitios web pueden diseñarse fácilmente para incitar a las personas a que superen sus sesgos. Si los alimentos saludables se colocan en lugares destacados, es probable que más personas los compren.

Una manera diferente de suprimir sesgos *ex ante* consiste en formar a los responsables de tomas de decisiones para que reconozcan sus sesgos y los superen. Algunas de estas intervenciones han recibido el nombre de

boosting; su objetivo es mejorar las capacidades de las personas (por ejemplo, enseñándoles rudimentos de estadística).[4]

Educar a las personas para que superen sus sesgos es una empresa honrosa, pero más ardua de lo que parece. Por supuesto, la educación es útil.[5] Por ejemplo, las personas que han asistido durante años a cursos de estadística avanzada son menos propensas a cometer errores de razonamiento estadístico. Sin embargo, enseñar a la gente a evitar los sesgos es difícil. Décadas de investigación han demostrado que los profesionales que han aprendido a evitar los sesgos en su área de especialización suelen tener dificultades para aplicar lo que han aprendido a otros campos. Los meteorólogos, por ejemplo, han aprendido a evitar el exceso de confianza en sus previsiones. Cuando anuncian un 70 por ciento de probabilidades de lluvia, llueve, por lo general, el 70 por ciento del tiempo. Sin embargo, pueden ser tan confiados como otras personas cuando se les hacen preguntas sobre temas de cultura general.[6] El reto de aprender a superar un sesgo es reconocer que un nuevo problema es similar a otro que hemos visto en otros ámbitos, y que un sesgo que hemos visto en un ámbito es probable que se materialice en otros.

Investigadores y educadores han tenido cierto éxito utilizando métodos de enseñanza no tradicionales para facilitar este reconocimiento. En un estudio realizado por Carey K. Morewedge, de la Universidad de Boston, y sus colegas, se utilizaron vídeos instructivos y «juegos serios». Los participantes aprendieron a reconocer los errores provocados por el sesgo de confirmación, el anclaje y otros sesgos psicológicos. Después de cada juego, recibían información sobre los errores que habían cometido y aprendían a no volver a caer en ellos. Los juegos (y, en menor medida, los vídeos) redujeron el número de errores que los participantes cometieron en un test inmediatamente posterior y, de nuevo, ocho semanas después, cuando les hicieron preguntas similares.[7] En un estudio independiente, Anne-Laure Sellier y sus colegas descubrieron que los estudiantes de MBA que habían jugado a un videojuego instructivo en el que aprendieron a superar el sesgo de confirmación aplicaron este aprendizaje para resolver un caso empresarial en otra clase. Lo hicieron a pesar de que no se les dijo que había una conexión entre los dos ejercicios.[8]

Tanto si corrigen los sesgos *ex post* como si evitan sus efectos mediante el *nudging* o el *boosting*, la mayoría de los métodos antisesgos tienen una cosa en común: se dirigen a un sesgo específico, que suponen está presente. Esta suposición, a menudo razonable, es a veces errónea.

Consideremos de nuevo el ejemplo de la planificación de proyectos. Se puede suponer razonablemente que el exceso de confianza afecta a los equipos planificadores de proyectos en general, pero no se puede asegurar que este sea el único sesgo (o el principal) que afecta a un equipo que planifica un proyecto concreto. Puede ser que el jefe del equipo haya tenido una mala experiencia con un proyecto similar y por eso ha aprendido a ser particularmente conservador a la hora de hacer estimaciones. Y así, el equipo exhibe el error opuesto al que pensaba que debía corregir. O tal vez el equipo elaboró su pronóstico por analogía con otro proyecto similar y se quedó anclado en el tiempo que tardó en completar ese proyecto. O tal vez el equipo planificador del proyecto, previendo que se añadiría un amortiguador a su estimación, se adelantó a ajustarla haciendo que su sugerencia sea aún más optimista de lo que verdaderamente cree.

Ahora, consideremos una decisión de inversión. Sin duda, el exceso de confianza en las perspectivas de inversión puede estar presente, pero otro poderoso sesgo, la aversión a las pérdidas, produce el efecto contrario, haciendo que los responsables de la decisión sean reacios a correr el riesgo de perder su inversión inicial. O pensemos en una empresa que asigna recursos a varios proyectos. Los responsables de las decisiones pueden ser tanto optimistas sobre el efecto de las nuevas iniciativas (de nuevo, exceso de confianza) como demasiado temerosos a la hora de desviar recursos de unidades existentes (un problema causado por el «sesgo del *statu quo*», que, como su nombre indica, es la preferencia por dejar las cosas como están).

Como ilustran estos ejemplos, es difícil saber con exactitud qué sesgos psicológicos están afectando a un juicio. En cualquier situación de cierta complejidad, pueden actuar múltiples sesgos psicológicos que conspiran para añadir errores en la misma dirección o que se compensan entre sí, con consecuencias impredecibles.

El resultado es que la supresión de sesgos *ex post* o *ex ante* —que, respectivamente, corrige o evita sesgos psicológicos específicos— es útil en

algunas situaciones. Estos métodos funcionan cuando la dirección que generalmente toma el error es conocida y se manifiesta como un claro sesgo estadístico. Los tipos de decisiones que se espera que estén muy sesgados serían los que se beneficien de las intervenciones antisesgos. Por ejemplo, la falacia de la planificación es un hallazgo lo bastante sólido como para justificar las intervenciones contra el exceso de confianza en la planificación.

El problema es que, en muchas situaciones, no se conoce de antemano la dirección probable del error. Entre estas se cuentan todas aquellas en las que el efecto de los sesgos psicológicos es variable entre los juzgadores y esencialmente impredecible, lo que origina ruido del sistema. Para reducir el error en circunstancias como estas, tenemos que lanzar una red más amplia para intentar detectar más de un sesgo psicológico a la vez.

EL OBSERVADOR DE LA DECISIÓN

Sugerimos emprender esta búsqueda de sesgos en tiempo real, no antes ni después de tomar la decisión. Por supuesto, las personas rara vez son conscientes de sus propios sesgos cuando son engañadas por ellos. Esta ausencia de conciencia es de por sí un sesgo conocido, el «sesgo del punto ciego».[9] Las personas suelen reconocer los sesgos con más facilidad en otros que en ellas mismas. Sugerimos que se podría formar a unos observadores para que detectasen en tiempo real signos que permitirían diagnosticar la presencia de uno o varios sesgos conocidos que están afectando a las decisiones o recomendaciones de una persona.

Para ilustrar cómo podría funcionar el proceso, imaginemos a un grupo que intenta hacer un juicio complejo y con consecuencias. El juicio puede ser de cualquier tipo: un Gobierno que decide sobre las posibles respuestas a una pandemia u otra crisis, unos médicos que discuten acerca del mejor tratamiento para un paciente con síntomas complejos o el consejo de administración de una empresa que decide sobre un movimiento estratégico importante. Ahora imaginemos a un «observador de la decisión», alguien que observa a este grupo y utiliza una lista de control para detectar si algún sesgo puede estar alejando al grupo del mejor juicio posible.

El papel de un observador de decisiones no es fácil de desempeñar y, sin duda, en algunas organizaciones no es práctico. Detectar los sesgos es inútil

si los responsables últimos de las decisiones no se comprometen a combatirlos. Los responsables de las decisiones tendrían que ser los que iniciasen el proceso de observación de la decisión y los que apoyasen el papel del observador de la decisión. Desde luego, no recomendamos a nadie que se autoproclame observador de decisiones. No ganará amigos ni influirá en nadie.

Con todo, los experimentos informales indican que se pueden hacer verdaderos progresos con este método, que es útil al menos en condiciones adecuadas, sobre todo cuando los líderes de una organización o de un equipo están realmente comprometidos con el esfuerzo, y cuando los observadores de las decisiones están bien elegidos (y no son susceptibles de tener sesgos propios importantes).

En estos casos, los observadores de las decisiones se clasifican en tres categorías. En algunas organizaciones, este papel puede desempeñarlo un supervisor. En lugar de supervisar solo el contenido de las propuestas presentadas por un equipo planificador de un proyecto, el supervisor también podría prestar atención al «proceso» de elaboración de estas y a la dinámica del equipo. Esto hace que el observador esté atento a los sesgos que puedan haber afectado al desarrollo de la propuesta.[10] Otras organizaciones pueden asignar a un miembro de cada equipo de trabajo para que sea el «destructor de sesgos» del equipo; este guardián del proceso de decisión recuerda a sus compañeros en tiempo real los sesgos que pueden inducirles a error. El inconveniente de este método es que el observador de la decisión se coloca en la posición de abogado del diablo dentro del equipo, y puede quedarse enseguida sin capital político. Por último, otras organizaciones pueden recurrir a un moderador externo, que presenta la ventaja de tener una perspectiva neutral (y las desventajas correspondientes en términos de conocimientos y costes internos).

Para ser eficaces, los observadores de decisiones necesitan cierta formación y herramientas. Una de estas últimas es una lista de control con los sesgos que se intentan detectar. Los argumentos para confiar en una lista de control son claros: las listas de control tienen un largo historial de mejora de las decisiones en contextos de alto riesgo, y son muy adecuadas para evitar la repetición de errores pasados.[11]

He aquí un ejemplo. En Estados Unidos, los organismos federales deben elaborar un análisis formal del impacto regulatorio antes de establecer

costosas normas destinadas a limpiar el aire o el agua, reducir las muertes en el lugar de trabajo, aumentar la seguridad alimentaria, responder a crisis de salud pública, reducir las emisiones de gases de efecto invernadero o aumentar la seguridad nacional. Un documento denso y técnico, con un nombre atroz (OMB Circular A-4) y con casi cincuenta páginas, establece los requisitos del análisis. Estos están claramente diseñados para contrarrestar los sesgos. Los organismos deben explicar por qué es necesaria la regulación, considerar alternativas más y menos estrictas, tener en cuenta tanto los costes como los beneficios, presentar la información de manera no sesgada y anticipar el futuro de forma apropiada. Sin embargo, en muchos organismos, los funcionarios de los estados no han cumplido con los requisitos de ese denso documento técnico. (Puede que ni siquiera lo hayan leído.) En respuesta, los funcionarios federales elaboraron una sencilla lista de control, de apenas página y media, para reducir el riesgo de que los organismos ignoren, o no cumplan, alguno de los principales requisitos.[12]

Para ilustrar el aspecto que podría tener una lista de control de los sesgos, hemos incluido una en el apéndice B.[13] Esta lista de control genérica es solo un ejemplo; cualquier observador de decisiones querrá sin duda desarrollar una que se adapte a las necesidades de la organización, tanto para demostrar su pertinencia como para favorecer su adopción.[14] Es importante destacar que una lista de control no es una lista exhaustiva de todos los sesgos que pueden afectar a una decisión; su objetivo es centrarse en los más frecuentes y con mayores consecuencias.

La observación de las decisiones con listas de control de sesgos adecuadas puede ayudar a limitar el efecto de los sesgos. Aunque hemos visto algunos resultados alentadores en intentos no convencionales a pequeña escala, no conocemos ninguna investigación sistemática de los efectos de este método o de los pros y los contras de las distintas formas posibles de implementarlo. Esperamos incentivar una mayor experimentación, tanto por parte de los profesionales como de los investigadores, en la práctica de la supresión de sesgos en tiempo real por parte de los observadores de decisiones.

El sesgo es un error que a menudo podemos detectar e incluso explicar, pues va en una determinada dirección; por eso un «empujón» (*nudge*) puede limitar los efectos perjudiciales de un sesgo, y por eso el intento de «potenciar» (*to boost*) el juicio puede combatir sesgos específicos. También es a menudo visible: por eso, un observador puede esperar diagnosticar sesgos en tiempo real mientras se toma una decisión.

El ruido, en cambio, es un error impredecible que no podemos ver o explicar con facilidad. Por eso lo descuidamos tan a menudo (incluso cuando provoca graves daños). Por esta razón, las estrategias de reducción del ruido son a la supresión de sesgos lo que las medidas higiénicas preventivas al tratamiento médico: el objetivo es prevenir una gama no especificada de posibles errores antes de que estos se produzcan.

A este enfoque de la reducción del ruido lo llamamos «higiene de la decisión». Cuando nos lavamos las manos, puede que no sepamos con precisión qué germen estamos evitando, solo sabemos que lavarse las manos es una buena prevención contra distintos gérmenes (sobre todo, pero no solo, durante una pandemia). Del mismo modo, seguir los principios de la higiene de las decisiones significa que adoptamos técnicas que reducen el ruido sin saber nunca qué errores subyacentes estamos contribuyendo a evitar.

La comparación con el lavado de manos es deliberada. Las medidas de higiene pueden ser tediosas. Sus beneficios no son visibles de inmediato; es posible que nunca sepamos qué problema han evitado. Y a la inversa: cuando surgen problemas, es posible que no sean atribuibles a un fallo específico en el cumplimiento a rajatabla de la higiene. Por estas razones, el lavado de manos es difícil de imponer, incluso entre los profesionales sanitarios, que son muy conscientes de su importancia.

Al igual que el lavado de manos y otras formas de prevención, la higiene de las decisiones es inestimable, pero ingrata. Corregir un sesgo bien identificado puede, al menos, darnos una sensación tangible de haber logrado algo. Pero los procedimientos que reducen el ruido no lo harán. Estadísticamente, evitarán muchos errores, pero nunca se sabrá «cuáles» son estos. El ruido es un enemigo invisible, e impedir el asalto de un enemigo invisible solo puede proporcionar una victoria invisible.

Sin embargo, si tenemos en cuenta el daño que puede provocar el ruido, esa victoria invisible merece, con todo, la pena. Los siguientes capítulos

presentan diferentes estrategias de higiene de la decisión utilizadas en múltiples campos, como la ciencia forense, la predicción, la medicina y los recursos humanos. En el capítulo 25, revisaremos estas estrategias y mostraremos cómo pueden combinarse en un método integrado para la reducción del ruido.

SOBRE LA SUPRESIÓN DE SESGOS E HIGIENE DE LA DECISIÓN

- «¿Sabe usted contra qué sesgo específico está luchando y en qué dirección afecta al resultado? Si no lo sabe, es probable que haya varios sesgos, y es difícil predecir cuál predominará.»
- «Antes de empezar a discutir esta decisión, vamos a designar a un observador de la decisión.»
- «Hemos mantenido una buena higiene en este proceso de decisión; es posible que nuestra decisión sea la mejor que podemos tomar.»

20
Secuenciación de la información en la ciencia forense

En marzo de 2004, una serie de bombas colocadas en trenes de cercanías mató a ciento noventa y tres personas e hirió a más de dos mil en Madrid. Una huella dactilar encontrada en una bolsa de plástico en el lugar del crimen se transmitió a través de Interpol a las fuerzas del orden de todo el mundo. Días después, el laboratorio de criminalística del FBI identificó la huella de forma concluyente: pertenecía a Brandon Mayfield, un ciudadano estadounidense residente en Oregón.

Mayfield parecía un sospechoso plausible. Antiguo oficial del ejército estadounidense, se había casado con una egipcia y se había convertido al islam. Como abogado, había representado a hombres acusados (y posteriormente condenados) de intentar viajar a Afganistán para unirse a los talibanes. Estaba en la lista de personas bajo vigilancia del FBI.

Mayfield fue sometido a vigilancia, su casa se intervino y se registró, y sus teléfonos fueron interceptados. Este examen detallado no arrojó ninguna información importante, y fue entonces cuando el FBI lo detuvo. Sin embargo, nunca se le acusó oficialmente. Mayfield no había salido del país en una década. Mientras estaba detenido, los investigadores españoles, que ya habían informado al FBI de que consideraban que no había coincidencia de la huella dactilar de Mayfield con la hallada en la bolsa de plástico, cotejaron esa huella con la de otro sospechoso.

Mayfield quedó en libertad a las dos semanas. El Gobierno estadounidense le pidió disculpas, le pagó una indemnización de dos millones de dólares y ordenó una amplia investigación sobre las causas del error. Su conclusión fue que «el error fue un error humano y no un fallo metodológico o tecnológico».[1]

Por fortuna, estos errores humanos son raros, pero resultan instructivos. ¿Cómo pudieron los mejores dactiloscopistas de Estados Unidos identificar erróneamente una huella dactilar como perteneciente a un hombre que nunca se había acercado al lugar del delito? Para averiguarlo, primero tenemos que entender cómo funciona el examen de huellas dactilares y

cómo se relaciona este con otros ejemplos de juicio profesional. Aprenderemos que el examen forense de huellas dactilares, que solemos considerar basado en una ciencia exacta, está, en realidad, sujeto a los sesgos psicológicos de los examinadores. Estos sesgos pueden generar más ruido y, por tanto, más error de lo que podríamos imaginar. También veremos cómo la comunidad científica forense está tomando medidas para abordar este problema mediante la puesta en práctica de una estrategia de higiene de las decisiones que puede aplicarse en todos los ámbitos: un control estricto del flujo de información utilizado para emitir juicios.

HUELLAS DACTILARES

Las huellas dactilares son las impresiones que dejan las yemas de nuestros dedos en las superficies que tocamos. Aunque hay ejemplos del uso de las huellas dactilares como aparentes marcas de identificación en la Antigüedad, la moderna toma de huellas dactilares se remonta a finales del siglo XIX, cuando Henry Faulds, un médico escocés, publicó el primer artículo científico que sugería el uso de las huellas dactilares como técnica de identificación.

En décadas posteriores, las huellas dactilares ganaron terreno como marcas de identificación en los registros criminales y sustituyeron de forma gradual a las técnicas de medición antropométrica desarrolladas por Alphonse Bertillon, un oficial de policía francés. El propio Bertillon codificó en 1912 un sistema formal para la comparación de huellas dactilares. Sir Francis Galton, a quien ya citamos como descubridor de la «sabiduría de la multitud», había desarrollado un sistema similar en el Reino Unido. (Sin embargo, no es de extrañar que estos padres fundadores sean raramente celebrados. Galton creía que las huellas dactilares serían una herramienta útil para clasificar a los individuos según su raza, y Bertillon, probablemente por prejuicios antisemitas, contribuyó con un testimonio pericial decisivo —y defectuoso— durante los procesos de 1894 y 1899 contra Alfred Dreyfus.)

Los agentes de policía pronto descubrieron que las huellas dactilares podían servir para algo más que para identificar a los delincuentes reincidentes. En 1892, Juan Vucetich, un policía argentino, fue el primero en comparar una huella dactilar oculta dejada en el lugar del delito con el

pulgar de un sospechoso. Desde entonces, la práctica de la recogida de «huellas ocultas» (las que deja su propietario en el lugar de un delito) y su comparación con las «huellas visibles» (las recogidas en condiciones controladas de individuos conocidos) ha sido la aplicación más decisiva de la dactiloscopia y ha constituido el tipo de prueba forense más utilizada.

Quien alguna vez haya visto un lector electrónico de huellas dactilares (como los utilizados en los controles de inmigración de muchos países), probablemente piense que la comparación de huellas dactilares es una tarea sencilla, mecánica y fácilmente automatizable. Sin embargo, comparar una huella oculta recogida en el lugar de un delito con una huella visible es un ejercicio mucho más delicado que comparar dos huellas limpias. Cuando se presionan los dedos con firmeza sobre un lector confeccionado a propósito para registrar la impresión de una huella dactilar, se produce una imagen nítida y estandarizada. Por el contrario, las huellas ocultas suelen ser parciales, poco claras, emborronadas o distorsionadas de algún modo; no proporcionan la misma cantidad y calidad de información que una huella recogida en un entorno controlado y específico. Las huellas ocultas suelen superponerse a otras huellas, ya sean de la misma persona o de otra, e incluyen suciedad y otros elementos presentes en la superficie. Decidir si coinciden con las huellas visibles de un sospechoso requiere un juicio experto. Y esta es la tarea de los dactiloscopistas.

Cuando reciben una huella oculta, los examinadores siguen habitualmente un proceso denominado ACE-V, que significa análisis, comparación, evaluación y verificación. En primer lugar, deben analizar la huella oculta para determinar si es lo suficientemente importante para ser comparada. Si lo es, la comparan con una huella visible. La comparación conduce a una evaluación, que puede producir una «identificación» (las huellas proceden de la misma persona), una «exclusión» (las huellas no proceden de la misma persona) o una decisión no concluyente. La decisión de que se ha logrado una identificación lleva al cuarto paso: la verificación por parte de otro examinador.

Durante décadas, la fiabilidad de este procedimiento fue incuestionable. Aunque se ha demostrado que los testimonios de los testigos oculares son peligrosamente dudosos, y que incluso las confesiones pueden ser falsas, las huellas dactilares se aceptaban —al menos hasta la llegada del análisis de ADN— como la prueba más creíble. Hasta 2002, las pruebas

dactiloscópicas nunca habían sido impugnadas con éxito en un tribunal estadounidense. El sitio web del FBI, por ejemplo, era en aquel entonces categórico: «Las huellas dactilares ofrecen un medio *infalible* de identificación personal».[2] En los rarísimos casos en que se producían errores, se achacaban a la incompetencia o al fraude.

Las pruebas dactiloscópicas no han sido cuestionadas durante tanto tiempo en parte por la dificultad de demostrar que son erróneas. A menudo se desconoce el valor real de un conjunto de huellas dactilares, es decir, la verdad sobre quién cometió realmente el delito. En el caso de Mayfield y un puñado de casos similares, el error fue particularmente enorme. Sin embargo, en general, si un sospechoso impugna las conclusiones del examinador, las pruebas dactilares se considerarán, por supuesto, más creíbles.

Ya hemos visto que no conocer el valor real no es ni algo excepcional ni un impedimento para medir el ruido. ¿Cuánto ruido hay en el análisis de las huellas dactilares? O, más exactamente, dado que los dactiloscopistas, a diferencia de los jueces o los aseguradores, no juzgan con un número, sino que emiten un juicio categórico, ¿con qué frecuencia discrepan y por qué? Esta pregunta es la que se propuso estudiar Itiel Dror, investigador de neurociencia cognitiva en el University College de Londres. Llevó a cabo lo que equivale a una serie de auditorías del ruido en un campo en el que se suponía que no había ningún problema de ruido.

RUIDO DE OCASIÓN EN EL ANÁLISIS DACTILOSCÓPICO

Puede parecer extraño que un científico cognitivo —un psicólogo— desafíe a los dactiloscopistas. Al fin y al cabo, como habrá visto el lector en series de televisión como *CSI. Investigación en la escena del crimen* y las series posteriores de la franquicia *CSI*, en ellas hay ciencia rigurosa con guantes de látex y microscopio. Pero Dror se dio cuenta de que examinar las huellas dactilares era sin duda una cuestión de juicio. Y, como neurocientífico cognitivo, razonó que dondequiera que haya juicio, hay ruido.

Para probar esta hipótesis, Dror se centró primero en el ruido de ocasión: la variabilidad entre los juicios de «los mismos» expertos que consideran dos veces la «misma» evidencia. Como dice Dror, «si los expertos no son

fiables en el sentido de que no son coherentes consigo mismos, entonces la base de sus juicios y su profesionalidad están en entredicho».[3]

Las huellas dactilares constituyen un banco de pruebas perfecto para una auditoría del ruido ocasional, ya que, a diferencia de los casos con los que se encuentra un médico o un juez, los pares de huellas no son fácilmente memorizables. Por supuesto, hay que dejar pasar un intervalo de tiempo adecuado para que los dactiloscopistas no recuerden las huellas. (En los estudios de Dror, algunos expertos valientes y de mente abierta estuvieron de acuerdo en que, «en cualquier momento de los siguientes cinco años», participarían en estudios sin que ellos lo supieran.) Además, el experimento debía realizarse en el transcurso del trabajo rutinario de los expertos, para que así no supieran que sus habilidades se estaban sometiendo a prueba. Si, en estas circunstancias, los juicios de los examinadores cambian de una prueba a otra, estamos en presencia de ruido de ocasión.

EL SESGO DE CONFIRMACIÓN FORENSE

En dos de sus primeros estudios, Dror añadió un giro importante. Al ver las huellas por segunda vez, algunos de los dactiloscopistas estuvieron expuestos a información nueva y sesgada sobre el caso. Por ejemplo, a los dactiloscopistas que antes habían encontrado coincidencias les dijeron esta vez cosas como «el sospechoso tiene una coartada», o «las pruebas de balística indican que no fue él». A otros, que primeramente habían llegado a la conclusión de que un sospechoso era inocente o de que las huellas no eran concluyentes, se les dijo la segunda vez que «el detective cree que el sospechoso es culpable», o que «los testigos presenciales lo identificaron», o que «confesó el delito». Dror consideró este experimento como un test de la «susceptibilidad al sesgo» de los expertos, porque la información contextual suministrada activaba un sesgo psicológico (un sesgo de confirmación) en una dirección determinada.

De hecho, los examinadores resultaron ser susceptibles al sesgo. Cuando los mismos examinadores consideraron las mismas huellas que habían visto anteriormente, pero esta vez con información sesgada, sus juicios cambiaron. En el primer estudio, cuatro de cada cinco expertos modificaron su decisión de identificación anterior cuando se les presentó información contextual sólida que sugería una exclusión.[4] En el segundo estudio, seis

expertos revisaron cuatro pares de huellas; la información sesgada provocó cambios en cuatro de las veinticuatro decisiones.[5] Sin duda, la mayoría de sus decisiones no cambiaron, pero, para este tipo de decisiones, un cambio en uno de cada seis puede calificarse de considerable. Estos hallazgos han sido reproducidos por otros investigadores.

Como era de esperar, los examinadores eran más propensos a cambiar de opinión cuando, para empezar, la decisión era difícil, cuando la información sesgada era sólida y cuando el cambio era de una decisión concluyente a otra no concluyente. No obstante, resulta preocupante que «los dactiloscopistas expertos tomaran decisiones basándose en el contexto, en lugar de hacerlo en la información real contenida en la huella».[6]

El efecto de la información sesgada no se limita a la conclusión del examinador (identificación, juicio no concluyente o exclusión). La información sesgada cambia realmente «lo que» el examinador percibe, además de «cómo» interpreta esa percepción. En otro estudio, Dror y sus colegas demostraron que los dactiloscopistas puestos en un contexto sesgado no ven literalmente las mismas cosas que los que no han sido expuestos a información sesgada.[7] Cuando la huella oculta va acompañada de una huella visible, los dactiloscopistas observan muchos menos detalles (llamados *minutiæ*) que cuando ven la huella oculta sola. Un estudio posterior e independiente confirmó esta conclusión y añadió que «no queda claro cómo se produce esto».[8]

Dror acuñó un término para el efecto de la información sesgada: el «sesgo de confirmación forense». Este sesgo se ha documentado desde entonces en otras técnicas forenses, como análisis de sangre, la investigación de incendios provocados, el análisis de restos óseos y la patología forense. Hasta el análisis de ADN —considerado en general como la nueva regla de oro en la ciencia forense— puede ser susceptible al sesgo de confirmación, al menos cuando los expertos deben evaluar mezclas complejas de ADN.[9]

La susceptibilidad al sesgo de confirmación entre los expertos forenses no solo es una preocupación teórica, porque, en realidad, no existen precauciones sistemáticas para garantizar que los expertos forenses no estén expuestos a información sesgada. Los examinadores suelen recibir esta información en las cartas de envío que acompañan a las pruebas que se les

presentan. Los dactiloscopistas también suelen estar en comunicación directa con la policía, los fiscales y otros examinadores.[10]

El sesgo de confirmación plantea otro problema. Una importante salvaguarda contra los errores, incorporada al procedimiento ACE-V, es la verificación independiente por parte de otro experto antes de confirmar una identificación. Sin embargo, lo más frecuente es que solo las identificaciones se verifiquen de forma independiente. El resultado es un fuerte riesgo de sesgo de confirmación, ya que el examinador verificador sabe que la conclusión inicial era una identificación.[11] Por tanto, la etapa de verificación no proporciona el beneficio que normalmente se espera de la suma de juicios independientes, porque las verificaciones no son, de hecho, independientes.

Una cascada de sesgos de confirmación parece haber actuado en el caso de Mayfield, en el que no dos, sino tres expertos del FBI coincidieron en la identificación errónea. Como se puso de manifiesto en la investigación posterior del error, el primer dactiloscopista quedó impresionado por «el poder de la correlación» del sistema automatizado que buscaba en los bancos de datos dactiloscópicos una posible coincidencia.[12] Aunque al parecer no se le dio a conocer ningún detalle biográfico de Mayfield, los resultados proporcionados por el sistema computarizado que llevó a cabo la búsqueda inicial, «junto con la presión de tener que trabajar en un caso de extrema relevancia», fueron suficientes para generar el sesgo de confirmación inicial. Una vez que el primer dactiloscopista realizó la identificación errónea, proseguía el informe, «los exámenes posteriores estuvieron contaminados». Como el primer dactiloscopista era un supervisor muy respetado, «cada vez era más difícil que otros dactiloscopistas de la agencia estuvieran en desacuerdo». El error inicial se reprodujo y amplificó, dando como resultado la casi certeza de que Mayfield era culpable. Resulta revelador que hasta un experto independiente muy respetado, nombrado por el tribunal para examinar las pruebas solicitadas por la defensa de Mayfield, coincidiera con el FBI en confirmar la identificación.[13]

El mismo fenómeno puede producirse en otras disciplinas forenses y entre ellas. La identificación de huellas ocultas tiene fama de ser una de las disciplinas forenses más objetivas. Si los dactiloscopistas pueden albergar sesgos, lo mismo puede ocurrir con los expertos en otros campos. Además,

si un experto en armas de fuego sabe que las huellas dactilares coinciden, este conocimiento puede sesgar también el juicio de dicho experto. Y si un odontólogo forense sabe que el análisis de ADN ha identificado a un sospechoso, es probable que ese experto sea menos propenso a sugerir que las marcas de mordedura no coinciden con la dentadura del sospechoso. Estos ejemplos trazan el espectro de las cascadas de sesgos: al igual que en las decisiones de grupo que describimos en el capítulo 8, un error inicial fruto del sesgo de confirmación se convierte en información sesgada que influye en un segundo experto, cuyo juicio sesga al de un tercero, y así sucesivamente.[14]

Una vez establecido que la información sesgada crea variabilidad, Dror y sus colegas descubrieron más pruebas de ruido de ocasión. Incluso cuando los dactiloscopistas no están expuestos a información sesgada, a veces cambian de opinión sobre un conjunto de huellas que han visto antes.[15] Como era de esperar, los cambios son menos frecuentes cuando no se suministra información sesgada, pero de todos modos se producen. Un estudio de 2012 encargado por el FBI reprodujo este descubrimiento a mayor escala pidiendo a setenta y dos examinadores que volvieran a mirar veinticinco pares de huellas que habían evaluado unos siete meses antes.[16] Con una amplia muestra de examinadores altamente cualificados, el estudio confirmó que los dactiloscopistas son a veces susceptibles al ruido de ocasión. Casi una de cada diez decisiones fue modificada. La mayoría de los cambios fueron a (o de) la categoría no concluyente, y ninguno dio lugar a falsas identificaciones. La implicación más preocupante del estudio es que algunas identificaciones de huellas dactilares que condujeron a condenas podrían haber sido juzgadas como no concluyentes en otro momento. Cuando los mismos dactiloscopistas examinan las mismas huellas, incluso si el contexto no está diseñado para crear sesgos, sino que se pretende que sea lo más estable posible, hay inconsistencia en sus decisiones.

ALGO DE RUIDO, PERO ¿CUÁNTO ERROR?

La cuestión práctica que plantean estos hallazgos es la posibilidad de los errores judiciales. No podemos ignorar las cuestiones relativas a la fiabilidad de los expertos que testifican ante los tribunales: la validez

requiere fiabilidad porque, sencillamente, es difícil estar de acuerdo con la realidad si no se está de acuerdo con uno mismo.

¿Cuántos errores exactamente son causados por una ciencia forense deficiente? Una revisión de trescientas cincuenta exculpaciones obtenidas por el Innocence Project, una organización sin ánimo de lucro que trabaja para anular las condenas injustas, concluyó que la aplicación errónea de la ciencia forense era una causa coadyuvante en el 45 por ciento de los casos.[17] Esta estadística suena mal, pero la cuestión que importa a los jueces y a los miembros de los jurados es diferente: para saber cuánta confianza deben otorgar al dactiloscopista que sube al estrado a testificar, necesitan saber cuál es la probabilidad de que los científicos forenses, incluidos los dactiloscopistas, cometan errores importantes.

El conjunto más sólido de respuestas a esta pregunta se encuentra en un informe del President's Council of Advisors on Science and Technology (PCAST, por sus siglas en inglés), un grupo asesor compuesto por los principales científicos e ingenieros del país que en 2016 llevó a cabo una revisión en profundidad de la ciencia forense en los tribunales penales.[18] El informe resume las pruebas disponibles sobre la validez del análisis de las huellas dactilares y, en particular, sobre la probabilidad de identificaciones erróneas (falsos positivos), como la de Mayfield.

Estas pruebas son sorprendentemente escasas y, como señala el PCAST, es «penoso» que los trabajos para elaborarlas no hayan comenzado hasta hace poco. Los datos más creíbles provienen del único estudio a gran escala que se ha publicado sobre la precisión de la identificación de huellas dactilares, que realizaron los propios científicos del FBI en 2011.[19] En el estudio participaron ciento sesenta y nueve dactiloscopistas, cada uno de los cuales comparó alrededor de cien pares de huellas dactilares ocultas y visibles. Su principal conclusión fue que se produjeron muy pocas identificaciones erróneas: la tasa de falsas identificaciones positivas fue de, aproximadamente, una de cada seiscientas.

Una tasa de error de una entre seiscientas es baja, pero, como señaló el informe, «*mucho más alta* de lo que el público en general (y, por extensión, la mayoría de los jurados) probablemente creería basándose en las afirmaciones de larga data sobre la exactitud del análisis de huellas dactilares».[20] Además, este estudio no contenía información contextual sesgada, y los dactiloscopistas sabían que estaban participando en una

prueba, lo que pudo haber hecho que el estudio subestimase los errores que se producen en el trabajo sobre casos reales. Un estudio posterior realizado en Florida llegó a cifras mucho más altas de falsos positivos.[21] Los diferentes descubrimientos en el material publicado sugieren que es necesario investigar más sobre la exactitud de las decisiones de los dactiloscopistas y sobre cómo se toman estas decisiones.

Con todo, un hallazgo tranquilizador que parece ser constante en todos los estudios es que los dactiloscopistas parecen pecar de precavidos. Su acierto no es completo, pero son conscientes de las consecuencias de sus juicios y tienen en cuenta el coste asimétrico de los posibles errores. Debido a la gran credibilidad de las huellas dactilares, una identificación errónea puede tener consecuencias fatales. Otros tipos de error son menos importantes. Por ejemplo, los expertos del FBI observan que «en la mayoría de los casos, una exclusión tiene las mismas implicaciones operativas que un resultado no concluyente».[22] En otras palabras, el hecho de que se encuentre una huella dactilar en el arma homicida es suficiente para condenar, pero la ausencia de esa huella no es suficiente para descartar a un sospechoso.

En consonancia con nuestra observación sobre la cautela de los dactiloscopistas, las pruebas sugieren que los expertos se lo piensan dos veces —o mucho más que dos veces— antes de tomar una decisión de identificación. En el estudio del FBI sobre la exactitud de las identificaciones, menos de un tercio de los pares «emparejados» (en los que la huella oculta y la visible son de la misma persona) fueron juzgados (con acierto) como identificaciones. Los dactiloscopistas también hacen muchas menos identificaciones positivas falsas que exclusiones negativas falsas.[23] Son susceptibles de sesgo, pero no por igual en ambas direcciones. Como señala Dror, «es más fácil inclinar a los expertos forenses hacia la conclusión de “no concluyente”, que no los compromete, que hacia la conclusión definitiva de “identificación”».[24]

Los dactiloscopistas están formados para considerar la identificación errónea como el pecado capital que hay que evitar cometer a toda costa. Para su credibilidad, actúan de acuerdo con este principio. Solo podemos esperar que su nivel de cautela haga que las identificaciones erróneas, como las del caso Mayfield y un puñado de otros casos de especial relevancia, sean extremadamente raras.

La observación de que en la ciencia forense hay ruido no debe entenderse como una crítica a los científicos forenses. No es más que una consecuencia de la observación que hemos hecho en repetidas ocasiones: dondequiera que haya juicio, hay ruido, y más del que se piensa. Una tarea como el análisis de las huellas dactilares parece objetiva, tanto que muchos de nosotros no la consideraríamos de forma espontánea una forma de juicio. Sin embargo, deja espacio a la incoherencia, el desacuerdo y, en ocasiones, el error. Por muy bajo que sea el índice de error en la identificación de huellas dactilares, no es cero, y, como señaló el PCAST, los jurados deben ser conscientes de ello.

El primer paso para reducir el ruido debe ser, por supuesto, reconocer su posibilidad. Este reconocimiento no resulta natural para los miembros de la comunidad de los dactiloscopistas, muchos de los cuales se mostraron muy escépticos en un principio con la auditoría del ruido de Dror. La idea de que un dactiloscopista pueda verse involuntariamente influido por información sobre el caso de que se ocupa irritó a muchos expertos. En una respuesta al estudio de Dror, el presidente de la asociación de dactiloscopistas escribió que «cualquier dactiloscopista que [...] se vea influido de cualquier manera en ese proceso de toma de decisiones [...] es tan inmaduro que debería buscar empleo en Disneylandia».[25] Un director de un importante laboratorio forense señaló que tener acceso a la información del caso —justo el tipo de información que podría sesgar al examinador— «proporciona cierta satisfacción personal que permite [a los examinadores] disfrutar de su trabajo *sin alterar realmente su juicio*».[26] Incluso el FBI, en su investigación interna del caso Mayfield, señaló que «los examinadores de huellas latentes realizan rutinariamente verificaciones en las que conocen los resultados de dactiloscopistas anteriores *y, sin embargo, esos resultados no influyen en las conclusiones del dactiloscopista*».[27] Estas observaciones equivalen claramente a una negación de la existencia del sesgo de confirmación.

Hasta cuando son conscientes del riesgo de sesgo, los forenses no son inmunes al sesgo del punto ciego: la tendencia a admitir la presencia de sesgos en otros, pero no en uno mismo. En una encuesta realizada a cuatrocientos forenses profesionales de veintiún países, el 71 por ciento

estaba de acuerdo en que «el sesgo cognitivo es un motivo de preocupación en el cuerpo de las ciencias forenses», pero solo el 26 por ciento pensaba que sus «propios juicios están influidos por el sesgo cognitivo».[28] En otras palabras, cerca de la mitad de estos profesionales forenses creen que los juicios de sus colegas son ruidosos, pero que los suyos no lo son. El ruido puede ser un problema invisible incluso para las personas cuyo trabajo es ver lo invisible.

SECUENCIAR LA INFORMACIÓN

Gracias a la perseverancia de Dror y sus colegas, las actitudes están cambiando lentamente, y un número creciente de laboratorios forenses ha empezado a tomar nuevas medidas para reducir los errores en sus análisis. Por ejemplo, el informe del PCAST elogió al laboratorio del FBI por haber rediseñado sus procedimientos para minimizar el riesgo del sesgo de confirmación.

Los pasos metodológicos necesarios son relativamente sencillos. Ilustran una estrategia de higiene de la decisión que tiene aplicación en muchos ámbitos: «secuenciar la información para limitar la formación de intuiciones prematuras». En cualquier juicio, una información es relevante y otra no. Más información no siempre es mejor, sobre todo si tiene el potencial de sesgar los juicios induciendo al juzgador a formarse una intuición prematura.

En este sentido, los nuevos procedimientos desplegados en los laboratorios forenses tienen como objetivo proteger la independencia de los juicios de los dactiloscopistas, dándoles solo la información que necesitan, cuando la necesitan. En otras palabras, el laboratorio los mantiene en la mayor oscuridad posible sobre el caso y revela la información solo de forma gradual. Por ello, el método que Dror y sus colegas codificaron se denomina «desenmascaramiento secuencial lineal».[29]

Dror tiene otra recomendación que ilustra la misma estrategia de higiene de las decisiones: los dactiloscopistas deben documentar sus juicios en cada paso. Deben documentar su análisis de una huella dactilar oculta «antes» de observar las huellas dactilares visibles para decidir si coinciden. Esta secuencia de pasos ayuda a los expertos a evitar el riesgo de que solo vean lo que buscan. Y deben registrar su juicio sobre las pruebas antes de tener

acceso a información contextual que pueda sesgarlo. Si cambian de opinión después de haber sido expuestos a información contextual, estos cambios, y la justificación de estos, deben ser documentados. Este requisito limita el riesgo de que una intuición precoz sesgue todo el proceso.

La misma lógica inspira una tercera recomendación, que es una parte importante de la higiene de las decisiones. Cuando un dactiloscopista diferente es llamado a verificar la identificación hecha por la primera persona, esa segunda persona no debe tener conocimiento del primer juicio.

La presencia de ruido en la ciencia forense es, por supuesto, preocupante por sus posibles consecuencias de vida o muerte. Sin embargo, también es reveladora. El hecho de que durante tanto tiempo hayamos ignorado por completo la posibilidad del error en la identificación de huellas dactilares demuestra que nuestra confianza en el juicio humano experto puede ser a veces exagerada, y que una auditoría del ruido puede revelar una cantidad inesperada de ruido. La capacidad de mitigar estas deficiencias mediante cambios relativamente sencillos en el proceso debería ser alentadora para todos aquellos que se preocupan por mejorar la calidad de las decisiones.

La principal estrategia de higiene de las decisiones que ilustra este caso —la información secuencial— tiene una amplia aplicabilidad como salvaguarda contra el ruido de ocasión. Como hemos señalado, el ruido de ocasión es el resultado de innumerables factores desencadenantes, como el estado de ánimo y hasta la temperatura exterior. No se puede esperar tener un control de todos estos desencadenantes, pero se puede intentar proteger a los juicios de los más obvios. Sabemos, por ejemplo, que los juicios pueden verse alterados por la indignación, el temor u otras emociones, por eso es una buena práctica revisar, si es posible, el propio juicio en distintos momentos, cuando es probable que los desencadenantes del ruido de ocasión sean diferentes.

Menos obvia es la posibilidad de que un juicio pueda verse alterado por otro factor desencadenante de ruido de ocasión: la información (incluso cuando es información precisa). Como en el ejemplo de los dactiloscopistas, en cuanto se sabe lo que piensan otros, el sesgo de confirmación puede hacer que se cree demasiado pronto una impresión general y que se ignore la información que la contradice. Los títulos de dos películas de Hitchcock resumen muy bien el asunto: quien se sienta

responsable de sus decisiones debe mantener «la sombra de una duda», no ser «el hombre que sabía demasiado».

SOBRE SECUENCIAR LA INFORMACIÓN

- «Dondequiera que haya juicio, hay ruido; esto incluye al examen de huellas dactilares.»
- «Tenemos más información sobre este caso, pero no digamos a los expertos todo lo que sabemos antes de que emitan su juicio, para no sesgarlos. Solo les diremos lo que necesitan saber.»
- «La segunda opinión no es independiente si la persona que la da sabe cuál fue la primera opinión. Y la tercera, aún menos; puede haber una cascada de sesgos.» ■ «Para luchar contra el ruido, primero hay que admitir que este existe.»

21
Selección y suma en los pronósticos

Muchos juicios implican un pronóstico. ¿Cuál será la tasa de desempleo en el próximo trimestre? ¿Cuántos coches eléctricos se venderán el próximo año? ¿Cuáles serán los efectos del cambio climático en 2050? ¿Cuánto tiempo se tardará en construir un nuevo edificio? ¿Cuáles serán los beneficios anuales de una determinada empresa? ¿Cómo se portará el nuevo empleado? ¿Cuál será el coste de una nueva normativa sobre contaminación atmosférica? ¿Quién ganará las elecciones? Las respuestas a estas preguntas tienen importantes consecuencias. De ellas dependen a menudo decisiones fundamentales de instituciones privadas y públicas.

Los analistas de los pronósticos —de cuándo estos fallan y por qué— distinguen claramente entre sesgo y ruido (también llamados «falta de coherencia» o «falta de fiabilidad»). Todo el mundo está de acuerdo en que, en algunos contextos, los pronósticos están sesgados. Por ejemplo, los organismos oficiales muestran un optimismo poco realista en sus previsiones presupuestarias.[1] Generalmente prevén un crecimiento económico excesivamente alto y un déficit excesivamente bajo. A efectos prácticos, poco importa que su optimismo irreal sea producto de un sesgo cognitivo o de consideraciones políticas.

Además, los analistas tienden a ser demasiado confiados: si se les pide que formulen sus pronósticos como intervalos de confianza en lugar de como estimaciones puntuales, tienden a elegir intervalos más estrechos de lo que deberían.[2] Por ejemplo, en una encuesta trimestral se pide a los directores financieros de empresas estadounidenses que calculen la rentabilidad anual del índice S&P 500 para el año siguiente. Los directores financieros dan dos cifras: una mínima, por debajo de la cual creen que hay una posibilidad entre diez de que se sitúen los beneficios reales, y otra máxima, por encima de la cual creen que los beneficios reales tienen una posibilidad entre diez de situarse. Por tanto, las dos cifras son los límites de un intervalo de confianza del 80 por ciento. Sin embargo, solo el 36 por ciento de las veces se quedan los beneficios obtenidos en ese intervalo. Los

directores financieros confían demasiado en la precisión de sus previsiones.[3]

Los analistas también son ruidosos. Un texto de referencia, *Principles of Forecasting*, de Jon Scott Armstrong, señala que, incluso entre los expertos, «la falta de fiabilidad es una fuente de error en los juicios que pronostican».[4] De hecho, el ruido es una de las principales fuentes de error. El ruido de ocasión es común; los analistas no siempre están de acuerdo consigo mismos. El ruido entre personas también está generalizado; los analistas no están de acuerdo entre sí, incluso si son especialistas. Si se pide a los profesores de Derecho que predigan las sentencias del Tribunal Supremo, se encontrará una gran cantidad de ruido.[5] Si se pide a los especialistas que calculen los beneficios anuales de la regulación de la contaminación atmosférica, se hallará una enorme variabilidad, con rangos de, por ejemplo, tres mil a nueve mil millones de dólares.[6] Si se pide a un grupo de economistas que hagan pronósticos sobre el desempleo y el crecimiento, también se descubrirá una gran variabilidad. Ya hemos visto muchos ejemplos de predicciones ruidosas, y la investigación sobre predicciones halla muchos más.[7]

MEJORA DE LOS PRONÓSTICOS

La investigación también ofrece sugerencias para reducir el ruido y el sesgo. No las examinaremos aquí de forma exhaustiva, pero nos centraremos en dos estrategias de reducción del ruido que tienen una amplia aplicabilidad. Una consiste en una aplicación del principio que mencionamos en el capítulo 18: la selección de mejores juzgadores produce mejores juicios. La otra es una de las estrategias de higiene de la decisión que tiene una aplicación más universal: el agregado de múltiples estimaciones independientes.

La manera más fácil de sumar varios pronósticos es hacer un promedio de estos. El promedio está matemáticamente garantizado para reducir el ruido: concretamente, lo divide por la raíz cuadrada del número de juicios promediados. Esto significa que, si hace una media de cien juicios, el ruido se reducirá en un 90 por ciento, y si hace una media de cuatrocientos juicios, lo reducirá en un 95 por ciento; esto es, lo eliminará. Esta ley

estadística es la que fundamenta el método de la sabiduría de la multitud analizado en el capítulo 7.

Dado que el acto de promediar no reduce el sesgo, su efecto sobre el error total depende de las proporciones de sesgo y ruido que contenga. Por eso, la sabiduría de la multitud funciona mejor cuando los juicios son independientes y, por tanto, es menos probable que contengan sesgos compartidos. Empíricamente, hay muchas pruebas que indican que promediar múltiples pronósticos aumenta en gran medida la exactitud, por ejemplo, en el pronóstico «de consenso» de los analistas económicos de bolsa.[8] En lo que respecta al pronóstico de ventas, la predicción meteorológica y la predicción económica, la media no ponderada de un grupo de analistas supera la mayoría de las predicciones individuales, y en ocasiones a todas ellas.[9] Promediar los pronósticos obtenidos por diferentes métodos tiene el mismo efecto: en un análisis de treinta comparaciones empíricas en diferentes campos, los pronósticos combinados redujeron los errores en una media del 12,5 por ciento.[10]

Hacer una media aritmética no es la única forma de sumar pronósticos. Una estrategia de la «selección de la multitud», que selecciona a los mejores juzgadores según el acierto de sus juicios recientes y promedia los juicios de un pequeño número de juzgadores (cinco, por ejemplo), puede ser tan efectiva como la media aritmética. También es más fácil para los responsables de la toma de decisiones que respetan la experiencia entender y adoptar una estrategia que no se base solo en la suma, sino también en la selección.[11]

Un método para producir un conjunto de pronósticos es utilizar los «mercados de predicciones», en los que los individuos apuestan sobre los resultados probables y, por tanto, se hallan incentivados para hacer predicciones correctas. Se ha comprobado que, la mayoría de las veces, los mercados de predicciones funcionan muy bien, en el sentido de que, si el precio del mercado de predicciones sugiere que los eventos tienen, digamos, un 70 por ciento de probabilidades de suceder, estos suceden alrededor del 70 por ciento de las veces.[12] Muchas empresas de diferentes sectores se han servido de los mercados de predicciones para agregar distintas perspectivas.[13]

Otro proceso formal para sumar diferentes perspectivas se conoce como el método Delphi. En su forma clásica, este implica numerosas rondas,

durante las cuales los participantes presentan estimaciones (o dan votos) a un moderador y mantienen su anonimato entre sí. En cada nueva ronda, los participantes justifican sus estimaciones y responden a las razones dadas por otros (siempre de forma anónima). El proceso impulsa la convergencia de las estimaciones (y a veces las obliga a hacerlo exigiendo que los nuevos juicios se sitúen dentro de un rango específico de la distribución de los juicios de la ronda anterior). El método se beneficia tanto de la suma como del aprendizaje social.[14]

El método Delphi ha funcionado bien en muchas situaciones, pero puede ser difícil de poner en práctica.[15] Una versión más sencilla, el «mini-Delphi», puede utilizarse en una sola reunión. También se denomina método de «estimar-hablar-estimar», y requiere que los participantes hagan primero estimaciones separadas (y calladas), que luego las expliquen y justifiquen, y que finalmente realicen una nueva estimación en respuesta a las estimaciones y explicaciones de los demás. El juicio de consenso es la media de las estimaciones individuales obtenidas en esa segunda ronda.[16]

EL GOOD JUDGMENT PROJECT

Algunos de los trabajos más innovadores sobre la calidad de la predicción, que van mucho más allá de lo que hemos investigado hasta ahora, comenzaron en 2011, cuando tres destacados científicos del comportamiento fundaron el Good Judgment Project. Philip Tetlock (a quien conocimos en el capítulo 11 cuando hablamos de su evaluación de las predicciones a largo plazo de acontecimientos políticos), su esposa Barbara Mellers y Don Moore se unieron para mejorar nuestra comprensión de las predicciones y, en particular, de la razón por la que algunas personas son buenas haciéndolas.

El Good Judgment Project se inició con el reclutamiento de decenas de miles de voluntarios; no especialistas ni expertos, sino personas corrientes de muchos ámbitos de la vida social. Se les pidió que respondieran a cientos de preguntas como, por ejemplo, las siguientes:

- ¿Detonará Corea del Norte un artefacto nuclear antes de que acabe el año?
- ¿Se anexionará Rusia oficialmente más territorio ucraniano en los próximos tres meses?
- ¿Se convertirá India o Brasil en miembro permanente del Consejo de Seguridad de la ONU en los próximos dos años?

• ¿Abandonará algún país la eurozona el próximo año?

Como muestran estos ejemplos, el proyecto se centró en grandes cuestiones relativas a acontecimientos mundiales. Es importante destacar que los intentos de responder a estas preguntas plantean muchos de los problemas de las predicciones más triviales. Si un abogado se pregunta si un cliente ganará en los tribunales, o si un estudio de televisión se pregunta si un programa propuesto será un gran éxito, hay que tener alguna capacidad para hacer pronósticos. Tetlock y sus colegas querían saber si algunas personas son particularmente buenas a la hora de pronosticar. También querían saber si esa capacidad podía enseñarse o, al menos, mejorarse.

Para entender los principales descubrimientos, debemos explicar algunos aspectos clave del método adoptado por Tetlock y su equipo para evaluar a los analistas. En primer lugar, utilizaron un gran número de predicciones —no una o unas pocas— en las que el éxito o el fracaso podría deberse a la suerte. Quien predice que su equipo deportivo favorito ganará el próximo partido y lo consigue, no es forzosamente un buen analista. Quizá siempre prediga que su equipo favorito va a ganar: si esa es su estrategia, y solo gana la mitad de las veces, su capacidad de predicción no es especialmente impresionante. Para reducir el papel de la suerte, los investigadores examinaron cómo les fue a los participantes, de media, en numerosos pronósticos.

En segundo lugar, los investigadores pidieron a los participantes que hicieran sus predicciones en términos de probabilidades de que un acontecimiento ocurriera, en lugar de un binario «ocurrirá» o «no ocurrirá». Para mucha gente, la predicción significa esto último: tomar partido en un sentido o en otro. Sin embargo, dada nuestra ignorancia objetiva de los acontecimientos futuros, es mucho mejor formular predicciones probabilísticas. Si alguien dijo en 2016: «Hillary Clinton tiene un 70 por ciento de probabilidades de ser elegida presidenta», no fue necesariamente un mal analista. Las cosas de las que se dice de manera adecuada que tienen un 70 por ciento de probabilidad no ocurrirán el 30 por ciento de las veces. Para saber si los pronósticos son buenos, debemos preguntarnos si sus cálculos de probabilidad se ajustan a la realidad. Supongamos que una analista particular llamada Margaret dice que quinientos acontecimientos

diferentes tienen un 60 por ciento de probabilidad. Si trescientos de ellos se producen, podemos concluir que la confianza de Margaret está bien «calibrada». Una buena calibración es un requisito para una buena predicción.

En tercer lugar, como perfeccionamiento añadido, Tetlock y sus colegas no se limitaron a pedir a sus analistas que hicieran «una» estimación de probabilidad sobre si un acontecimiento ocurriría en, digamos, doce meses. Dieron a los participantes la oportunidad de revisar continuamente sus predicciones a la luz de información nueva. Supongamos que alguien hubiera estimado, allá por 2016, que el Reino Unido solo tenía un 30 por ciento de probabilidades de abandonar la Unión Europea antes de finales de 2019. A medida que salían nuevos sondeos que indicaban que el voto a favor de la «salida» estaba ganando terreno, probablemente habría revisado su previsión al alza. Cuando se conoció el resultado del referéndum, aún no se sabía si el Reino Unido abandonaría la Unión Europea en ese plazo, pero sin duda parecía mucho más probable. (El Brexit se produjo en sentido estricto en 2020.)

Con cada nueva pieza de información, Tetlock y sus colegas permitían a los pronosticadores actualizar sus predicciones. A efectos de puntuación, cada una de estas actualizaciones se tomaba como una nueva predicción. De ese modo, los participantes en el Good Judgment Project se veían incentivados a seguir las noticias y actualizar continuamente sus predicciones. Este método refleja lo que se espera de los analistas en las empresas y los gobiernos, que también deben actualizar sus predicciones con frecuencia en función de la nueva información, a pesar del riesgo de que los critiquen por cambiar de opinión. (Una conocida respuesta a esta crítica, atribuida a veces a John Maynard Keynes, es: «Cuando los hechos cambian, yo cambio de opinión. ¿Y “usted”?».)

En cuarto lugar, para puntuar la capacidad de los analistas, el Good Judgment Project utilizó un sistema desarrollado por Glenn W. Brier en 1950. Las «puntuaciones Brier», como se conocen, miden la distancia entre lo que se pronostica y lo que realmente sucede.

Las puntuaciones Brier son una forma inteligente de obviar un problema general asociado a las predicciones probabilísticas: las razones que puedan tener los analistas para minimizar riesgos y no adoptar nunca una actitud más osada. Pensemos de nuevo en Margaret, a la que describimos como una

analista bien calibrada porque dio a quinientos acontecimientos un 60 por ciento de probabilidad y trescientos de esos acontecimientos se produjeron. Este resultado puede no ser tan impresionante como parece. Si Margaret es una meteoróloga que predice «siempre» un 60 por ciento de probabilidades de lluvia y hay trescientos días lluviosos de quinientos, las predicciones de Margaret están bien calibradas, pero también son inútiles en la práctica. Margaret no está diciendo más que, por si acaso, conviene llevar el paraguas todos los días. Comparémosla con Nicholas, que predice una probabilidad de lluvia del cien por cien en los trescientos días en que lloverá, y una probabilidad de lluvia del 0 por ciento en los doscientos días sin lluvia. Nicholas tiene la misma calibración ideal que Margaret: cuando cualquiera de los analistas predice que el X por ciento de los días será lluvioso, la lluvia cae justo el X por ciento de las veces. Sin embargo, las predicciones de Nicholas son mucho más valiosas: en lugar de minimizar riesgos, está dispuesto a decirnos si debemos sacar el paraguas. Técnicamente, se dice que Nicholas tiene una alta «resolución», además de una buena calibración.

Las puntuaciones Brier recompensan tanto la buena calibración como la buena resolución. Para obtener una buena puntuación, no solo hay que acertar de media (es decir, estar bien calibrado), sino también estar dispuesto a diferenciar entre predicciones (es decir, tener una alta resolución). Las puntuaciones Brier se basan en la lógica de los errores cuadráticos medios (ECM), y las puntuaciones más bajas son mejores: una puntuación de 0 sería perfecta.

Así que, ahora que sabemos cómo se los puntuó, ¿qué tal lo hicieron los voluntarios del Good Judgment Project? Uno de los principales descubrimientos fue que la inmensa mayoría de los voluntarios lo hicieron mal, pero alrededor del 2 por ciento sobresalieron. Como ya se mencionó, Tetlock llama a estos últimos «superanalistas». No eran infalibles, pero sus predicciones eran mucho mejores que el azar. Sorprendentemente, un funcionario del Gobierno dijo que el grupo lo hizo muchísimo «mejor que la media de los analistas pertenecientes a la comunidad de los servicios de inteligencia que podían leer interceptaciones y otros datos secretos».[17] Merece la pena detenerse en esta comparación. Los analistas de la comunidad de inteligencia están formados para hacer predicciones exactas;

no son aficionados. Además, tienen acceso a información clasificada. Y, sin embargo, no lo hacen tan bien como los superanalistas.

¿Qué hace que los superanalistas sean tan buenos? De acuerdo con nuestro argumento en el capítulo 18, podríamos especular de manera razonable que son extraordinariamente inteligentes. Esta especulación no es incorrecta. En las pruebas de las calificaciones medias, los superanalistas obtienen mejores resultados que la media de los voluntarios del Good Judgment Project (y la media de los voluntarios está significativamente por encima de la media nacional). Sin embargo, la diferencia no es tan grande, y muchos voluntarios que obtienen resultados extraordinarios en los test de inteligencia no pueden calificarse de superanalistas. Aparte de la inteligencia general, podríamos esperar de forma razonable que los superanalistas sean excepcionalmente buenos con los números. Y lo son. Pero su verdadera ventaja no es su talento para las matemáticas, sino su facilidad para pensar de forma analítica y probabilística.

Hay que tener en cuenta la voluntad y la capacidad de los superanalistas para estructurar y desglosar los problemas. En lugar de formarse un juicio holístico sobre una gran cuestión geopolítica (si una nación abandonará la Unión Europea, si estallará una guerra en un lugar concreto, si un funcionario público será asesinado), desglosan sus partes. Se preguntan: «¿Qué hace falta para que la respuesta sea afirmativa? ¿Qué hace falta para que la respuesta sea negativa?». En lugar de ofrecer un instinto visceral o una especie de premonición global, preguntan e intentan responder a un surtido de preguntas secundarias.

Los superanalistas también sobresalen por su visión externa y se preocupan mucho por las tasas-base. Como se explicó para el problema de Gambardi en el capítulo 13, antes de que el lector se centre en los detalles del perfil de Gambardi, le puede ser útil conocer la probabilidad de que el director ejecutivo medio sea despedido o renuncie en los próximos dos años. Los superanalistas buscan de manera sistemática las tasas-base. Ante la pregunta de si en el próximo año se producirá un enfrentamiento armado entre China y Vietnam por una disputa fronteriza, los superpronósticos no se centrarán solo o sin más en si China y Vietnam se llevan bien en el

presente. Pueden tener una intuición al respecto sobre la base de las noticias y los análisis que han leído. Sin embargo, saben que su intuición sobre un acontecimiento no suele ser una buena guía. Por eso comienzan buscando una tasa-base: preguntan con qué frecuencia las disputas fronterizas anteriores han devenido en enfrentamientos armados. Si estos últimos son poco frecuentes, los superanalistas empezarán por incorporar ese hecho, y solo entonces pasarán a analizar los detalles de la situación entre China y Vietnam.

En resumen, lo que distingue a los superanalistas no es su inteligencia, sino «cómo» la utilizan. Las habilidades que emplean reflejan el tipo de estilo cognitivo que describimos en el capítulo 18 como el capaz de producir mejores juicios, en particular un alto grado de «mentalidad activamente abierta». Recordemos el test de pensamiento activamente abierto: incluye afirmaciones como «La gente debería tener en cuenta las pruebas que van en contra de sus creencias» y «Es más útil prestar atención a las personas que no están de acuerdo con uno que a las que sí lo están». Está claro que las personas que obtienen una puntuación alta en este test no vacilan en actualizar sus juicios (sin reaccionar de forma exagerada) cuando disponen de nueva información.

Para caracterizar el estilo de pensamiento de los superanalistas, Tetlock utiliza la expresión «beta perpetua», un término utilizado por los programadores informáticos para referirse a un programa que no está destinado a ser lanzado en una versión final, sino que se utiliza, analiza y mejora sin cesar. Tetlock descubre que «el mejor factor de predicción para ascender al rango de los superanalistas es la beta perpetua, el grado de compromiso con la actualización de las creencias y la superación personal». Y añade: «Lo que los hace tan buenos no es tanto lo que son como lo que hacen: el duro trabajo de investigación, la cuidadosa reflexión con autocrítica, la recopilación y síntesis de otras perspectivas, los juicios granulares y la incesante actualización».[18] Los atrae un particular ciclo de pensamiento: «probar, fallar, analizar, ajustar, volver a probar».[19]

RUIDO Y SESGO EN LOS PRONÓSTICOS

Llegados a este punto, uno podría pensar que es posible preparar a los individuos para que sean superanalistas o, al menos, se parezcan más a

ellos. De hecho, Tetlock y sus colaboradores han trabajado para hacer exactamente eso. Sus esfuerzos por lograrlo deben considerarse la segunda etapa para entender por qué los superanalistas funcionan tan bien y cómo hacer que aún funcionen mejor.

En un relevante estudio, Tetlock y su equipo formaron tres grupos de manera aleatoria con los analistas habituales (no los súper), en los que probaron el efecto de diferentes intervenciones sobre la calidad de juicios posteriores. Estas últimas ejemplifican tres de las estrategias que hemos expuesto para mejorar los juicios:

1. *Formación*. Varios pronosticadores completaron un tutorial diseñado para mejorar sus habilidades en el que se les enseñaba el razonamiento probabilístico. En el tutorial, los analistas aprendieron a identificar varios sesgos (incluida la ignorancia de la tasa-base, el exceso de confianza y el sesgo de confirmación) y a ser conscientes de la importancia de promediar múltiples predicciones de distintas fuentes y de considerar clases de referencia.
2. *Trabajo en equipo (una forma de suma)*. Se pidió a algunos analistas que trabajaran en equipos en los que pudieran ver y debatir las predicciones de otros. El trabajo en equipo podría aumentar la exactitud al animar a los analistas a tratar con argumentos opuestos y ser activamente abiertos.
3. *Selección*. Se puntuó la exactitud de todos los analistas y, al cabo de un año, se otorgó a un 2 por ciento de ellos, los más destacados, el título de superanalistas, y, al año siguiente, se les dio la oportunidad de trabajar juntos en equipos de élite.

Resultó que las tres intervenciones funcionaron, en el sentido de que mejoraron las puntuaciones Brier de los participantes. La formación marcó una diferencia, el trabajo en equipo otra mayor y la selección tuvo un efecto aún mayor.

Este importante descubrimiento confirma el valor de la suma de juicios y la selección de buenos juzgadores. Pero aquí no acaba la historia. Armado con datos sobre los efectos de cada intervención, Ville Satopää, que colaboró con Tetlock y Mellers, desarrolló una compleja técnica estadística para averiguar cómo mejoraba exactamente cada intervención las predicciones.[20] En principio, razonó, hay tres razones principales por las que algunos analistas pueden resultar mejores o peores que otros:

1. Pueden ser más hábiles a la hora de encontrar y analizar datos del entorno que sean relevantes para la predicción que tienen que hacer. Esta explicación señala la importancia de la información.
2. Algunos analistas pueden tener una tendencia general a errar en un lado particular del valor real de una predicción. Si, entre cientos de predicciones, alguien sobrestima o subestima de manera sistemática la probabilidad de que se produzcan determinados cambios en el *statu quo*, se puede decir que muestra una forma de sesgo a favor del cambio o de la estabilidad.

3. Algunos analistas pueden ser menos susceptibles al ruido (o a los errores aleatorios). En la predicción, como en cualquier juicio, el ruido puede tener muchos desencadenantes. Los analistas pueden reaccionar de forma exagerada ante una noticia concreta (esto es un ejemplo de lo que hemos llamado «ruido de patrón»), pueden estar sujetos al ruido de ocasión o pueden ser ruidosos en su uso de la escala de probabilidad. Todos estos errores (y muchos más) son impredecibles en su magnitud y dirección.

Satopää, Tetlock, Mellers y su colega Marat Salikhov llamaron a su modelo el modelo BIN (*bias, information, noise*) de predicción. Se propusieron medir el grado en que cada uno de los tres componentes, considerados en cada una de las tres intervenciones, contribuía a la mejora del rendimiento.

Su respuesta fue sencilla: las tres intervenciones funcionaban principalmente reduciendo el ruido. Según los investigadores, «siempre que una intervención aumentaba la exactitud, lo hacía sobre todo suprimiendo los errores aleatorios en el juicio. Curiosamente, la intención original de la intervención era reducir el sesgo».[21]

Dado que la formación se diseñó para reducir los sesgos, un analista que no alcanzaba el grado «súper» habría predicho que la reducción del sesgo sería el principal efecto del entrenamiento. Sin embargo, lo que hizo la preparación fue reducir el ruido. La sorpresa se explica fácilmente. La preparación de Tetlock está pensada para combatir los sesgos «psicológicos». Y, como ya sabemos, el efecto de los sesgos psicológicos no siempre es un sesgo estadístico. Cuando afectan de maneras distintas a diferentes individuos en diferentes juicios, los sesgos psicológicos producen ruido. Este es sin duda el caso, ya que los acontecimientos que se pronostican son muy variados. Los mismos sesgos pueden hacer que un analista reaccione de manera exagerada o leve dependiendo del tema. No debemos esperar que produzcan un sesgo «estadístico», definido como la tendencia general de un analista a creer que los acontecimientos ocurrirán o no. Por ello, la formación de los analistas para que luchen contra sus sesgos psicológicos funciona (reduciendo el ruido).

El trabajo en equipo tuvo un efecto comparativamente grande en la reducción del ruido, pero también mejoró de forma significativa la capacidad de los equipos para extraer información. Este resultado es coherente con la lógica de la suma: varios cerebros que trabajan juntos son mejores para encontrar información que uno solo. Si María y Juan trabajan

juntos, y María ha detectado señales que a Juan se le han escapado, su predicción conjunta será mejor. Cuando trabajan en grupo, los superanalistas parecen capaces de evitar los peligros de la polarización de grupo y las cascadas de información. Para conseguirlo, se comunican sus datos y conocimientos y, con su mentalidad activa, aprovechan al máximo la información combinada. Satopää y sus colegas explican así esta ventaja: «El trabajo en equipo, a diferencia de la preparación [...], permite a los analistas aprovechar la información».[22]

La selección tuvo el mayor efecto total. Parte de la mejora proviene de un mejor uso de la información. Los superanalistas son mejores que los demás a la hora de encontrar información relevante (posiblemente porque son más inteligentes, están más motivados y tienen más experiencia en hacer este tipo de pronósticos que el participante medio). Pero el principal efecto de la selección es, de nuevo, reducir el ruido. Los superanalistas son menos ruidosos que los analistas habituales, o incluso que los equipos preparados. Este descubrimiento también fue una sorpresa para Satopää y los demás investigadores: «Los "superanalistas" podrían deber su éxito más a una disciplina superior a la hora de impedir el error de medición que a una lectura incisiva de las noticias» que otros no pueden hacer.[23]

DÓNDE FUNCIONAN LA SELECCIÓN Y LA SUMA

El éxito del proyecto de superanálisis pone de manifiesto el valor de dos estrategias de la higiene de las decisiones: la «selección» (los superanalistas son, sí, superiores) y la «suma» (cuando trabajan en equipo, obtienen mejores resultados). Las dos estrategias son ampliamente aplicables a muchos juicios. Siempre que sea posible, hay que intentar combinar las estrategias formando equipos de juzgadores (por ejemplo, analistas, profesionales de la inversión, contratadores) seleccionados por ser buenos en lo que hacen y complementarios entre sí.

Hasta ahora hemos considerado la mejora de la precisión que se consigue promediando múltiples juicios independientes, como en los experimentos de sabiduría de la multitud. La suma de las estimaciones de los juzgadores de mayor validez mejorará aún más la exactitud. Sin embargo, se puede ganar más en exactitud combinando juicios que son al mismo tiempo independientes y complementarios.[24] Imaginemos que cuatro personas

son testigos de un delito: es esencial, por supuesto, asegurarse de que no se influyan mutuamente. Si, además, lo han visto desde cuatro ángulos diferentes, la calidad de la información que proporcionen será mucho mejor.

La tarea de reunir a un equipo de profesionales para que emitan juicios de forma conjunta se asemeja a la de reunir una serie de test para predecir el futuro rendimiento de candidatos en la universidad o en el trabajo. La herramienta estándar para esta tarea es la regresión múltiple (introducida en el capítulo 9). Esta funciona seleccionando variables en sucesión. Se selecciona primero el test que mejor predice el resultado. Sin embargo, el siguiente test que se incluye no es necesariamente el segundo más válido, sino el que «añade» el mayor poder predictivo al primer test al proporcionar predicciones que son válidas y no superficiales con respecto al primero. Por ejemplo, supongamos que tenemos dos test de aptitud mental que se correlacionan en .50 y .45 con el rendimiento futuro, y un test de personalidad que solo se correlaciona en .30 con el rendimiento, pero que no se correlaciona con los test de aptitud. La solución óptima es elegir primero el test de aptitud más válido y luego el test de personalidad, que aporta más información nueva.

Del mismo modo, si se está formando un equipo de juzgadores, hay que elegir primero, por supuesto, al mejor juzgador. Pero la siguiente elección puede ser un individuo medianamente válido que aporte alguna habilidad nueva en lugar de un juzgador más válido que sea muy similar al primero. Un equipo seleccionado de esta manera será superior, porque la validez de los juicios sumados aumenta más rápidamente cuando los juicios no están correlacionados entre sí que cuando son superfluos. El ruido de patrón será relativamente alto en un equipo de este tipo, ya que los juicios individuales de cada caso serán diferentes. Paradójicamente, el promedio de este grupo ruidoso será más exacto que el de uno unánime.

Es preciso hacer una importante advertencia. Aparte de la diversidad, la suma solo puede reducir el ruido si los juicios son en verdad independientes. Como ha puesto de manifiesto nuestro análisis del ruido en los grupos, la deliberación en grupo a menudo añade más errores de sesgo de los que elimina de ruido. Las organizaciones que quieran aprovechar el poder de la diversidad deben aceptar los desacuerdos que surgirán cuando los miembros del equipo hagan sus juicios de forma independiente. La

estrategia de higiene de las decisiones más fácil, barata y de amplia aplicación es la de obtener y sumar juicios independientes y variados.

SOBRE SELECCIÓN Y SUMA

- «Obtengamos el promedio de cuatro juicios independientes; así se garantiza la reducción del ruido a la mitad.»
- «Deberíamos esforzarnos por estar en “beta perpetua”, como los superanalistas.»
- «Antes de debatir sobre esta situación, ¿cuál es la tasa-base pertinente?» ■ «Tenemos un buen equipo, pero ¿cómo podemos tener una mayor variedad de opiniones?»

22
Directrices en la medicina

Hace algunos años, un buen amigo nuestro (llamémosle Paul) fue diagnosticado de hipertensión por su médico de atención primaria (lo llamaremos doctor Jones). Este aconsejó a Paul que probara con un determinado medicamento. Le recetó un diurético, pero no surtió efecto; la presión arterial de Paul seguía siendo alta. Unas semanas después, el doctor Jones le propuso un segundo medicamento, un bloqueador de los canales de calcio. Su efecto fue bastante leve.

Estos resultados desconcertaron al doctor Jones. Después de tres meses de consultas semanales, la tensión de Paul había descendido ligeramente, pero seguía siendo demasiado alta. No tenía claro cuáles serían los siguientes pasos. Paul estaba ansioso y el doctor Jones preocupado, sobre todo porque Paul era un hombre relativamente joven y con buena salud. El doctor Jones contempló la posibilidad de probar con un tercer medicamento.

En ese momento, Paul se trasladó a otra ciudad, donde consultó a un nuevo médico de atención primaria (lo llamaremos doctor Smith). Paul le contó al doctor Smith la historia de sus continuas luchas contra la hipertensión. El doctor Smith respondió inmediatamente: «Compre un kit para medirse la tensión y utilícelo en casa. No creo que tenga usted la tensión alta en absoluto. Probablemente solo tenga el síndrome de la bata blanca: ¡su tensión sube en las consultas médicas!».

Paul hizo lo que el nuevo médico le dijo y, en efecto, su tensión era normal en casa. Desde entonces ha sido normal (y un mes después de que el doctor Smith le hablara del síndrome de la bata blanca, también fue normal en las consultas médicas).

Una de las principales tareas de los médicos es hacer diagnósticos, es decir, decidir si un paciente padece algún tipo de enfermedad y, si así es, identificarla. El diagnóstico suele requerir algún tipo de juicio. Para muchas afecciones, el diagnóstico es rutinario, y en gran medida mecánico, y existen normas y procedimientos para minimizar el ruido. Para un médico

suele ser fácil determinar si alguien tiene un hombro dislocado o un dedo del pie roto. Algo parecido puede decirse de problemas más técnicos. Cuantificar la degeneración de los tendones entraña poco ruido.[1] Cuando los patólogos evalúan las biopsias con aguja gruesa de las lesiones mamarias, sus evaluaciones son relativamente sencillas, con escaso ruido.[2]

Es importante destacar que algunos diagnósticos no implican juicio alguno. La atención sanitaria a menudo progresa eliminando el elemento de juicio (cambiando el juicio por el cálculo). En el caso de la faringitis estreptocócica, el médico comenzará con un test rápido de antígenos sobre una muestra de la garganta del paciente. En poco tiempo, el test puede detectar el estreptococo. (Sin este rápido resultado con el antígeno, y hasta cierto punto incluso con él, hay ruido en el diagnóstico de la faringitis estreptocócica.[3]) Si alguien tiene en ayunas un nivel de glucosa en la sangre de 126 miligramos por decilitro o superior, o una HbA1c (una medida media de la glucosa en la sangre durante los tres meses anteriores) de al menos 6,5, se considera que padece diabetes.[4] Durante las primeras etapas de la pandemia de la COVID-19, algunos médicos realizaron inicialmente diagnósticos como resultado de juicios hechos tras considerar los síntomas; a medida que la pandemia avanzaba, los test se hicieron mucho más comunes, y los test hicieron innecesario el juicio.

Mucha gente sabe que, cuando los médicos emiten su juicio, pueden ser ruidosos y pueden equivocarse; una práctica habitual es aconsejar a los pacientes que pidan una segunda opinión. En algunos hospitales, la segunda opinión es incluso obligatoria.[5] Cuando la segunda opinión diverge de la primera, tenemos ruido (aunque, por supuesto, puede no estar claro qué médico tiene razón). Algunos pacientes (incluido Paul) se han quedado sorprendidos al ver lo mucho que difiere la segunda opinión de la primera. Sin embargo, la sorpresa no es la existencia de ruido en la profesión médica; es la magnitud de este.

Nuestro objetivo en este capítulo es analizar este hecho y describir algunos de los métodos de reducción del ruido que utiliza el cuerpo médico. Nos centraremos en una estrategia de higiene de las decisiones: la elaboración de directrices para el diagnóstico. Somos muy conscientes de que se podría escribir todo un libro sobre el ruido en la medicina y las distintas medidas que médicos, enfermeras y hospitales han tomado para

ponerle remedio. El ruido en el ámbito particular de la medicina no se limita al ruido en los juicios de diagnóstico, que es nuestro tema aquí. Los tratamientos también pueden ser ruidosos, y un gran número de publicaciones abordan también este tema. Si un paciente tiene un problema cardiaco, las opiniones de los médicos sobre el mejor tratamiento varían increíblemente, tanto si se trata de la medicación adecuada como del tipo de intervención quirúrgica correcta o de la conveniencia de operarse. El Dartmouth Atlas Project se ha dedicado, durante más de veinte años, a documentar «las flagrantes variaciones en la distribución y el uso de los recursos médicos en Estados Unidos».[6] A conclusiones similares se ha llegado en numerosos países.[7] Sin embargo, para nuestros fines, bastará con una breve investigación del ruido en los juicios de diagnóstico.

UNA PANORÁMICA

Las publicaciones existentes sobre el ruido en la medicina son innúmeras. Aunque gran parte de ellas son empíricas y comprueban la presencia de ruido, buena parte de ellas son también prescriptivas. Los profesionales de la sanidad siguen buscando estrategias de reducción del ruido, que adoptan muchas formas y son una mina de valiosas ideas que merece la pena tener en cuenta en muchos campos.

Cuando hay ruido, un médico puede sin duda tener razón y otro estar sin duda equivocado (y puede sufrir algún tipo de sesgo). Naturalmente, las habilidades son de gran importancia. Un estudio sobre los diagnósticos de neumonía hechos por radiólogos, por ejemplo, reveló un ruido significativo.[8] Gran parte del ruido se debe a diferencias de habilidad. Más concretamente, «la variación en habilidad puede explicar el 44 por ciento de la variación en las decisiones de diagnóstico», lo que indica que «las políticas que mejoran la habilidad funcionan mejor que las directrices de decisión uniformes». En este caso, como en otros, la formación y la selección son evidentemente cruciales para la reducción del error y la eliminación del ruido y del sesgo.[9]

En algunas especialidades, como la radiología y la patología, los médicos son muy conscientes de la presencia de ruido. Los radiólogos, por ejemplo, llaman a la variación diagnóstica su «talón de Aquiles».[10] No está claro si el ruido en los campos de la radiología y la patología recibe una atención

especial, porque de hecho hay más ruido en estos campos que en otros, o simplemente porque el ruido es más fácil de documentar en ellos. Sospechamos que la facilidad para documentarlo puede ser más importante. Los sencillos test de ruido (y a veces de error) son más fáciles de hacer en radiología. Por ejemplo, se puede volver a los escaneos o a las placas para reevaluar una valoración anterior.

En la medicina, el ruido entre personas, o la «fiabilidad entre evaluadores», suele medirse usando la «estadística kappa».[11] Cuanto mayor sea la kappa, menor será el ruido. Un valor kappa de 1 refleja un acuerdo total; un valor de 0 refleja exactamente la concordancia que cabría esperar entre monos que lanzan dardos a una lista de posibles diagnósticos. En algunos ámbitos del diagnóstico médico, la fiabilidad medida por este coeficiente ha resultado ser «leve» o «pobre», lo que significa que el ruido es muy alto. A menudo se la encuentra «regular», algo que, por supuesto, es mejor, pero que también indica un ruido significativo. En cuanto a la importante cuestión de qué interacciones farmacológicas son clínicamente significativas, los médicos generalistas, al revisar cien interacciones farmacológicas seleccionadas al azar, mostraron un «escaso acuerdo».[12] A los profanos y a muchos médicos, el diagnóstico de las distintas fases de la insuficiencia renal puede parecerles relativamente sencillo. Sin embargo, los nefrólogos solo muestran un «acuerdo entre leve y moderado» en sus juicios sobre el significado de las pruebas estándar utilizadas en la evaluación de los pacientes con insuficiencia renal.[13]

En cuanto a la cuestión de si una lesión mamaria es cancerosa, un estudio solo encontró entre los patólogos un acuerdo «regular».[14] En el diagnóstico de las lesiones proliferativas de la mama, el acuerdo fue de nuevo solo «regular».[15] El acuerdo también fue «regular» cuando los médicos evaluaron resonancias magnéticas para determinar el grado de estenosis espinal.[16] Merece la pena detenerse en estas conclusiones. Hemos dicho que, en algunos ámbitos, el nivel de ruido en la medicina es muy bajo. Sin embargo, en algunas áreas bastante técnicas, los médicos están lejos de evitar el ruido. Que a un paciente se le diagnostique una enfermedad grave, como un cáncer, puede depender de una especie de lotería, determinada por el médico concreto que lo examine.

Consideremos otros resultados de las diferentes publicaciones, extraídos de áreas en las que el volumen de ruido parece particularmente notable. No

describimos estos resultados para hacer declaraciones autorizadas sobre el estado actual de la práctica médica, que sigue evolucionando y mejorando (en algunos casos rápidamente), sino para transmitir una idea general de la omnipresencia del ruido, tanto en el pasado relativamente reciente como en el presente.

1. En Estados Unidos, los infartos de miocardio son la principal causa de muerte tanto en hombres como en mujeres.[17] Los angiogramas coronarios, uno de los métodos más utilizados para diagnosticar el infarto de miocardio, evalúan el grado de obstrucción de las arterias del corazón tanto en casos agudos como no agudos. En los casos no agudos, cuando un paciente presenta un dolor torácico recurrente, suele recibir un tratamiento —como la colocación de un *stent*— si se encuentra que una o más arterias mayores están obstruidas en un 70 por ciento. Sin embargo, se ha documentado cierto grado de variabilidad en la interpretación de los angiogramas, lo que puede dar lugar a procedimientos innecesarios. Un primer estudio descubrió que el 31 por ciento[18] de las veces, los médicos que evaluaban los angiogramas no estaban de acuerdo en si un vaso principal estaba obstruido en más del 70 por ciento. A pesar de que los cardiólogos son conscientes de la posible variabilidad en la interpretación de los angiogramas, y a pesar de los continuos esfuerzos y medidas correctoras, el problema aún no se ha resuelto.
2. La endometriosis es una afección en la que el tejido endometrial, que normalmente recubre el interior del útero, crece fuera de este. Esta afección puede ser dolorosa y crear problemas de fertilidad. A menudo se diagnostica mediante una laparoscopia, una técnica consistente en introducir quirúrgicamente una pequeña cámara en el cuerpo. Se mostraron a ciento ocho cirujanos ginecológicos vídeos digitales de laparoscopias realizadas a tres pacientes, dos de las cuales tenían endometriosis de distinta gravedad y una no. Se pidió a los cirujanos que juzgaran el número y la localización de las lesiones endometriales. No coincidieron en nada, con correlaciones leves tanto en el número como en la ubicación.[19]
3. La tuberculosis es una de las enfermedades más extendidas y mortales en todo el mundo. Solo en 2016 se infectaron más de diez millones de personas, y casi dos millones murieron. Un método muy utilizado para detectar la tuberculosis es la radiografía de tórax, que permite examinar los pulmones en busca del espacio vacío producto de la infección por el bacilo de la tuberculosis. La variabilidad en el diagnóstico de la tuberculosis está bien documentada desde hace casi setenta y cinco años. A pesar de las mejoras introducidas a lo largo de décadas, los estudios han seguido encontrando una variabilidad significativa en el diagnóstico de tuberculosis, con una concordancia «moderada» o solo «regular» entre evaluadores.[20] También existe una variabilidad en los diagnósticos de tuberculosis entre radiólogos en diferentes países.[21]
4. Cuando unos patólogos analizaron lesiones cutáneas para detectar la presencia de un melanoma —la forma más peligrosa de cáncer de piel—, solo hubo un acuerdo «moderado». Los ocho patólogos que revisaron cada caso fueron unánimes o solo mostraron un desacuerdo en el 62 por ciento de las ocasiones.[22] Otro estudio realizado en un centro de oncología descubrió que el diagnóstico correcto del melanoma era solo del 64 por ciento, lo que significa que los médicos diagnosticaban erróneamente melanomas en una de cada tres lesiones.[23] Un tercer estudio reveló que los dermatólogos de la Universidad de Nueva York no diagnosticaron el melanoma a partir de biopsias de piel en el 36 por ciento de las ocasiones. Los autores del estudio concluyen que «el fracaso clínico en el diagnóstico correcto del melanoma tiene graves implicaciones para la supervivencia de los pacientes con esa enfermedad potencialmente mortal».[24]

5. Existe una variabilidad en los juicios de los radiólogos con respecto al cáncer de mama a partir de las mamografías de cribado. Un amplio estudio descubrió que el rango de falsos negativos entre diferentes radiólogos variaba desde el 0 por ciento (el radiólogo acertaba todas las veces) hasta más del 50 por ciento (el radiólogo consideraba incorrectamente que la mamografía era normal más de la mitad de las veces). Asimismo, las tasas de falsos positivos oscilaron entre menos del 1por ciento y el 64 por ciento (lo que significa que casi dos tercios de las veces el radiólogo sostuvo que la mamografía mostraba cáncer cuando este no estaba presente).[25] Los falsos negativos y los falsos positivos de diferentes radiólogos revelan la existencia de ruido.

Estos casos de ruido interpersonal dominan la investigación existente, pero también se ha descubierto ruido de ocasión. A veces, los radiólogos dan una opinión diferente al volver a evaluar la misma imagen y, por tanto, no están de acuerdo consigo mismos (aunque con menos frecuencia que con otros).[26] Al evaluar el grado de obstrucción en los angiogramas, veintidós médicos discreparon consigo mismos entre el 63 y el 92 por ciento de las veces.[27] En áreas que implican criterios vagos y juicios complejos, la «fiabilidad intraevaluador», como se la denomina, puede ser escasa.[28]

Estos estudios no ofrecen una explicación clara de este ruido de ocasión. Sin embargo, otro estudio, que no tiene que ver con el diagnóstico, identifica una sencilla fuente de ruido de ocasión en la medicina (un hallazgo que merece la pena tener en cuenta tanto para los pacientes como para los médicos).[29] Resulta que los médicos son significativamente más propensos a solicitar pruebas de detección de cáncer a primera hora de la mañana que a última hora de la tarde. En una amplia muestra, las tasas de solicitud de pruebas para detectar cáncer de mama y de colon fueron más altas a las ocho de la mañana en un 63,7 por ciento. Disminuyeron a lo largo de la mañana hasta el 48,7 por ciento a las once horas. Aumentaron hasta el 56,2 por ciento al mediodía, y luego disminuyeron hasta el 47,8 por ciento a las diecisiete horas. De ello se deduce que los pacientes con citas a horas más tardías tenían menos probabilidades de someterse a las pruebas de detección de cáncer recomendadas por las directrices.

¿Cómo explicar estos descubrimientos? Una posible respuesta es que los médicos se retrasan casi inevitablemente en la clínica después de ver a pacientes con problemas médicos complejos que requieren más tiempo del habitual (veinte minutos). Ya hemos mencionado el papel del estrés y el cansancio como desencadenantes del ruido de ocasión (véase el capítulo 7), y estos elementos parecen estar aquí presentes. Para cumplir con sus horarios, algunos médicos se saltan conversaciones sobre medidas

preventivas. Otra ilustración del papel del cansancio entre los doctores es la menor frecuencia del conveniente lavado de manos hacia el final de los turnos en el hospital.[30] (El lavado de manos también acaba siendo ruidoso.)

MÉDICOS MENOS RUIDOSOS: EL VALOR DE LAS DIRECTRICES

Sería una importante contribución (no solo a la medicina, sino también al conocimiento humano) ofrecer una relación exhaustiva de la existencia y la magnitud del ruido en el contexto de diferentes problemas médicos.[31] No tenemos constancia de que exista tal relación; esperamos que aparezca con el tiempo. Sin embargo, actualmente, los descubrimientos existentes proporcionan algunas pistas.

En un extremo, el diagnóstico de algunos problemas y enfermedades es esencialmente mecánico y no deja espacio para el juicio. En otros casos, el diagnóstico no es mecánico, sino sencillo; es muy probable que cualquier persona con formación médica llegue a la misma conclusión. En otros casos, un grado de especialización —entre, por ejemplo, los especialistas en cáncer de pulmón— será suficiente para garantizar que el ruido, aunque exista, sea mínimo. En el otro extremo, algunos casos ofrecen un gran margen para el juicio, y los criterios pertinentes para el diagnóstico son tan abiertos que el ruido será considerable y difícil de reducir. Como veremos, este es un caso frecuente en la psiquiatría.

¿Qué podría funcionar para reducir el ruido en la medicina? Como hemos mencionado, la formación puede aumentar la habilidad, y la habilidad desde luego ayuda.[32] También lo hace la suma de múltiples juicios de expertos (segundas opiniones, etcétera).[33] Los algoritmos ofrecen una vía muy prometedora, y los médicos ya están empleando algoritmos de aprendizaje profundo e inteligencia artificial (IA) para reducir el ruido. Por ejemplo, estos algoritmos se han utilizado para detectar metástasis en los ganglios linfáticos de las mujeres con cáncer de mama. Se ha comprobado que los mejores de ellos son superiores al mejor patólogo y, por supuesto, los algoritmos no son ruidosos.[34] Los algoritmos de aprendizaje profundo también se han usado, con considerable éxito, para la detección de padecimientos oculares asociados a la diabetes.[35] Y la IA funciona ahora al menos tan bien como los radiólogos en la detección del cáncer en las

mamografías; nuevos avances en la IA probablemente demuestren su superioridad.[36]

Es probable que la profesión médica recurra cada vez más a los algoritmos en el futuro. Los algoritmos prometen reducir tanto los sesgos como el ruido y, por consiguiente, salvar vidas y ahorrar dinero. Sin embargo, aquí haremos hincapié en las directrices del juicio humano, porque el ámbito de la medicina ilustra de una forma bastante útil los buenos, incluso excelentes, resultados que dan en algunas aplicaciones y los resultados más desiguales que dan en otras.

Quizá el ejemplo más célebre de una directriz de diagnóstico sea la puntuación de Apgar, desarrollada en 1952 por la anestesista obstétrica Virginia Apgar. Evaluar si un recién nacido está en peligro solía ser una cuestión de juicio clínico para médicos y matronas. Sin embargo, la puntuación de Apgar les dio toda una directriz. El evaluador mide el color, la frecuencia cardiaca, los reflejos, el tono muscular y el esfuerzo respiratorio del recién nacido, lo que a veces se resume en un acrónimo del apellido Apgar: *appearance* (apariencia; color de la piel), *pulse* (pulso; frecuencia cardiaca), *grimace* (mueca; reflejos), *activity* (tono muscular) y *respiration* (frecuencia y esfuerzo respiratorios). En el test de Apgar, cada una de estas cinco medidas recibe una puntuación de 0, 1 o 2. La puntuación total más alta posible es 10, la cual es rara. Una puntuación de 7 o más se considera un indicador de buena salud (tabla 3).[37]

Nótese que la frecuencia cardiaca es el único componente estrictamente numérico de la puntuación, y que todos los demás elementos implican un elemento de juicio. Sin embargo, como el juicio se descompone en elementos individuales, cada uno de los cuales es sencillo de evaluar, resulta poco probable que los asistentes al parto, aun con un grado modesto de formación, estén en franco desacuerdo (y, por tanto, la puntuación de Apgar produce poco ruido).[38]

La puntuación de Apgar ejemplifica cómo funcionan las directrices y por qué reducen el ruido. A diferencia de las reglas o los algoritmos, las directrices no eliminan la necesidad de juicio: la decisión no es un cálculo directo. Sigue siendo posible el desacuerdo sobre cada uno de los componentes y, por tanto, sobre la conclusión final. Sin embargo, las directrices consiguen reducir el ruido, porque descomponen una decisión

compleja en una serie de subjuicios más sencillos sobre aspectos ya definidos.

Categoría	Puntos asignados
Apariencia (color de la piel)	0: Todo el cuerpo está amoratado o pálido 1: Buen color en el cuerpo, pero manos o pies amoratados 2: Completamente rosado o de color normal
Pulso (frecuencia cardiaca)	0: Sin pulso 1: <100 pulsaciones por minuto 2: >100 pulsaciones por minuto
Mueca (reflejos)	0: Sin respuesta a la estimulación de las vías respiratorias 1: Mueca durante la estimulación 2: Mueca y tos o estornudo durante la estimulación
Actividad (tono muscular)	0: Renquera 1: Cierta flexión de brazos y piernas 2: Movimiento activo
Respiración (frecuencia y esfuerzo respiratorios)	0: Sin respiración 1: Llanto débil (gemidos, gruñidos) 2: Llanto fuerte

TABLA 3. Directrices de Apgar con sus puntuaciones

Las ventajas de este enfoque son evidentes cuando consideramos el problema desde el punto de vista de los modelos de predicción sencillos que analizamos en el capítulo 9. Un médico que emite un juicio sobre la salud de un recién nacido se basa en varias pistas predictivas. Podría existir ruido de ocasión: un día, pero no otro, o con un estado de ánimo determinado, pero no con otro, podría prestar atención a predictores relativamente poco importantes o ignorar los importantes. La puntuación de Apgar centra al profesional sanitario en los cinco aspectos que se sabe empíricamente que son importantes. A continuación, la puntuación proporciona una descripción clara de cómo evaluar cada pista, lo que en gran medida simplifica cada juicio fundado en las pistas, y, por tanto, reduce su ruido. Por último, la puntuación de Apgar especifica cómo ponderar los predictores mecánicamente para producir el juicio global requerido, mientras que los humanos diferirían en la ponderación que asignan a los

indicios. La concentración en los predictores relevantes, la simplificación del modelo de predicción y la suma mecánica reducen el ruido.

Se han utilizado enfoques similares en muchos ámbitos médicos. Un ejemplo es la puntuación de Centor para guiar el diagnóstico de la faringitis estreptocócica. Se da al paciente un punto por cada uno de los siguientes síntomas o signos (cuyos términos, al igual que en la puntuación de Apgar, constituyen un acrónimo parcial del apellido de Robert Centor, que resumió con sus colegas esta directriz por primera vez): ausencia de tos (*cough*), presencia de exudados (*exudates*; manchas blancas en la parte posterior de la garganta), sensibilidad o inflamación en los ganglios (*nodes*) linfáticos del cuello y una temperatura (*temperature*) superior a 38 grados centígrados. Dependiendo del número de puntos que se asignen al paciente, puede recomendarse un frotis de garganta para diagnosticar la faringitis estreptocócica. La evaluación y la puntuación son relativamente sencillas gracias a esta escala, que ha reducido eficazmente el número de personas que se someten a pruebas y tratamientos innecesarios de la faringitis estreptocócica.[39]

También se han establecido directrices de este tipo para el diagnóstico del cáncer de mama con el Breast Imaging Reporting and Data System (BI-RADS), que reduce el ruido en las interpretaciones de las mamografías. Un estudio descubrió que el BI-RADS aumentó la concordancia entre evaluadores de las mamografías, lo que demostraba que las directrices pueden ser eficaces para reducir el ruido en un área donde la variabilidad ha sido significativa.[40] También en patología se han llevado a cabo muchos intentos con éxito para utilizar directrices con el mismo fin.[41]

EL DEPRIMENTE CASO DE LA PSIQUIATRÍA

En lo referente al ruido, la psiquiatría es un caso extremo. Cuando diagnostican a un mismo paciente utilizando los mismos criterios de diagnóstico, los psiquiatras suelen discrepar entre sí. Por esta razón, la reducción del ruido ha sido una de las principales prioridades de la comunidad psiquiátrica al menos desde la década de 1940.[42] Y, como veremos, a pesar de que se perfeccionan continuamente, las directrices solo han prestado una modesta ayuda para reducir el ruido.

Un estudio realizado en 1964, en el que participaron noventa y un pacientes y diez psiquiatras experimentados, descubrió que la probabilidad de que hubiera un acuerdo entre dos opiniones era solo del 57 por ciento.[43] Otro de los primeros estudios, en el que participaron cuatrocientos veintiséis pacientes de hospitales estatales diagnosticados de forma independiente por dos psiquiatras, encontró un acuerdo de solo el 50 por ciento de las veces en su diagnóstico del tipo de enfermedad mental que padecían. Y otro estudio de la época, en el que participaron ciento cincuenta y tres pacientes ambulatorios, encontró un 54 por ciento de acuerdo. En estos estudios no se especificó la fuente del ruido. Sin embargo, es interesante observar que algunos psiquiatras se inclinaban por incluir a los pacientes en categorías diagnósticas específicas. Por ejemplo, algunos psiquiatras eran muy propensos a diagnosticar depresión, y otros, ansiedad.

Como pronto veremos, los niveles de ruido siguen siendo altos en psiquiatría. ¿Por qué? Los especialistas no dan una respuesta única y clara (lo que significa que las explicaciones del ruido son, ellas mismas, ruidosas). El nutrido conjunto de categorías de diagnóstico es, sin duda, un factor. Sin embargo, en un intento preliminar por responder a esa pregunta, los investigadores pidieron a un psiquiatra que entrevistara a un paciente, y luego a un segundo psiquiatra que realizara otra entrevista tras un breve periodo de descanso.[44] Los dos psiquiatras se reunieron después y, si no estaban de acuerdo, discutían por qué no lo estaban.

Una razón frecuente era la «inconstancia del médico»: diferentes escuelas de pensamiento, diferente formación, diferentes experiencias clínicas, diferentes estilos en las entrevistas. Mientras que un «médico con formación en desarrollo podía explicar la experiencia alucinatoria como parte de la experiencia postraumática de un abuso pasado», otro médico «con orientación biomédica podía explicar las mismas alucinaciones como parte de un proceso esquizofrénico».[45] Estas diferencias son ejemplos de ruido de patrón.

Sin embargo, más allá de las discrepancias entre los médicos, la principal razón del ruido era la «falta de una terminología adecuada». Estas observaciones y el descontento profesional generalizado con la terminología psiquiátrica motivaron la revisión de 1980 (la tercera edición) del *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III). El manual incluía por primera vez criterios claros y pormenorizados para

diagnosticar trastornos mentales, un primer paso hacia la introducción de directrices para los diagnósticos.

El DSM-III produjo un aumento espectacular de la investigación sobre el ruido en los diagnósticos. También demostró su utilidad para reducir el ruido.[46] Sin embargo, el manual estaba lejos de ser un completo éxito.[47] Incluso después de una importante revisión en el año 2000 de la cuarta edición (el DSM-IV, publicado originalmente en 1994), la investigación demostró que el nivel de ruido seguía siendo elevado.[48] Por un lado, Ahmed Aboraya y sus colegas concluyen que «se ha demostrado que el uso de criterios de diagnóstico para los trastornos psiquiátricos aumenta la fiabilidad de los diagnósticos psiquiátricos».[49] Por otro lado, sigue existiendo un riesgo serio de que «los ingresos de un paciente revelen múltiples diagnósticos para el mismo paciente».[50]

En 2013 salió otra versión del manual, el DSM-5.[51] La American Psychiatric Association esperaba que el DSM-5 redujera el ruido, pues esta nueva edición se basaba en criterios más objetivos y escalas más claras.[52] Sin embargo, los psiquiatras siguen acusando un ruido significativo.[53] Por ejemplo, Samuel Lieblich y sus colegas han descubierto que «a los psiquiatras les cuesta ponerse de acuerdo sobre quién tiene y quién no un trastorno depresivo grave».[54] Las pruebas de campo para el DSM-5 encontraron un «acuerdo mínimo», lo que «significa que, en las condiciones del estudio, psiquiatras especializados y altamente capacitados solo pudieron coincidir en que un paciente tiene depresión entre el 4 y el 15 por ciento de las veces».[55] Según algunos ensayos de campo, el DSM-5 en realidad empeoró las cosas, con mayor ruido «en todas las esferas principales, con algunos diagnósticos, como el trastorno mixto ansioso-depresivo [...] tan poco fiables que parecen inútiles en la práctica clínica».[56]

Al parecer, la principal razón del limitado éxito de las directrices es que, en psiquiatría, «los criterios diagnósticos de algunos trastornos siguen siendo vagos y difícilmente operativos».[57] Algunas directrices reducen el ruido al descomponer el juicio en criterios con los que se reduce el desacuerdo, pero en la medida en que estos son relativamente abiertos, el ruido sigue siendo probable. Teniendo en cuenta este punto, las propuestas más destacadas reclaman unas directrices de diagnóstico más

estandarizadas. Entre ellas, (1) la clarificación de los criterios diagnósticos, para alejarlos de las normas vagas; (2) la elaboración de «definiciones de referencia» para los síntomas y su nivel de gravedad sobre la base de la teoría de que, cuando «los médicos se ponen de acuerdo sobre la presencia o ausencia de síntomas, es más probable que se pongan de acuerdo sobre el diagnóstico», y (3) el servirse de entrevistas estructuradas a los pacientes, además de conversaciones abiertas. Una guía de entrevistas propuesta incluye veinticuatro preguntas de detección que permiten un diagnóstico más fiable de, por ejemplo, la ansiedad, la depresión y los trastornos alimentarios.[58]

Estos pasos se nos antojan prometedores, pero hasta qué punto conseguirían reducir el ruido es una cuestión abierta. En palabras de un observador, «la dependencia de los síntomas subjetivos del paciente, la interpretación de los síntomas por parte del médico y la ausencia de una medida objetiva (como lo es un análisis de sangre) echan las semillas de la falta de fiabilidad diagnóstica de los trastornos psiquiátricos».[59] En este sentido, la psiquiatría puede mostrarse especialmente resistente a los intentos de reducir el ruido.

Sobre esta cuestión en particular es demasiado pronto para hacer una predicción segura, aunque una cosa es clara: en la medicina en general, las directrices han tenido mucho éxito a la hora de reducir tanto el sesgo como el ruido; han ayudado a médicos, enfermeras y pacientes, y han ido mejorando considerablemente la salud pública. La profesión médica necesita aún más directrices.[60]

SOBRE DIRECTRICES EN LA MEDICINA

- «Entre los médicos, el nivel de ruido es mucho mayor de lo que habríamos sospechado. En el diagnóstico del cáncer y de las cardiopatías —y hasta en la interpretación de las radiografías—, los especialistas no se ponen a veces de acuerdo. Esto significa que el tratamiento que recibe un paciente puede ser producto de una lotería.»
- «Los médicos seguramente pensarán que toman las mismas decisiones sea lunes o viernes, o a primera hora de la mañana o a última hora de la tarde. Sin

embargo, resulta que lo que los médicos dicen y hacen bien podría depender de lo cansados que estén.»

- «Las directrices médicas pueden hacer que los médicos sean menos propensos a cometer errores a costa del paciente. Estas directrices también pueden ayudar al cuerpo médico en su conjunto, porque reducen la variabilidad.»

23
Definición de la escala en las calificaciones de rendimiento

Empecemos con un ejercicio. Escoja el lector a tres personas que conozca; pueden ser amigos o colegas. Valórelos en una escala de 1 a 5, donde 1 es lo más bajo y 5 lo más alto, en función de tres características: amabilidad, inteligencia y diligencia. Luego pida a alguien que los conozca bien —su cónyuge, su mejor amigo o su colega más cercano— que haga lo mismo con esas tres personas.

Es muy probable que, en algunas de las valoraciones, el lector y el otro evaluador tengan cifras diferentes. Si el lector (y el otro) está dispuesto, discuta las razones de las diferencias. Es posible que la respuesta se encuentre en su forma de utilizar la escala, lo que hemos llamado «ruido de nivel». Tal vez pensó que un 5 requiere algo realmente extraordinario, mientras que el otro evaluador pensó que simplemente requiere algo extraordinariamente bueno. O tal vez el lector difiera por tener distintos puntos de vista sobre las personas calificadas: su concepto de la amabilidad —y cómo definir exactamente esa virtud— podría ser diferente de la del otro calificador.

Ahora imagine el lector que para las tres personas que calificó está en juego un ascenso o una gratificación. Supongamos que él y el otro evaluador se dedican a calificar el rendimiento en una empresa que valora la amabilidad (o el compañerismo), la inteligencia y la diligencia. ¿Habría alguna diferencia entre sus valoraciones? ¿Serían tan altas como en el ejercicio anterior? ¿Incluso más altas? Sea cual fuere la respuesta a estas preguntas, es probable que las diferencias en las políticas y las escalas produzcan ruido. Y, de hecho, eso es lo que se observa de forma generalizada en las calificaciones de rendimiento en todos los entornos de las organizaciones.

En casi todas las grandes organizaciones, el rendimiento se evalúa formalmente de manera periódica. Los que son calificados no disfrutan de la experiencia. Como decía el titular de un periódico: «Un estudio revela que prácticamente todas las personas odian las valoraciones de su rendimiento».[1] Todo el mundo sabe también (creemos) que las evaluaciones de rendimiento están sujetas al sesgo y el ruido. Sin embargo, la mayoría desconoce lo ruidosas que son.

En un mundo ideal, la evaluación del rendimiento de los individuos no sería una tarea del juicio; los hechos objetivos bastarían para determinar lo bien que trabajan. Pero la mayoría de las organizaciones modernas tienen poco en común con la fábrica de alfileres de Adam Smith, en la que cada trabajador tiene un rendimiento medible. ¿Cuál sería ese rendimiento para un director financiero o para un director de investigación? Los trabajadores del conocimiento actuales equilibran numerosos objetivos, a veces contradictorios. Centrarse en solo uno de ellos podría generar evaluaciones erróneas y tener efectos incentivadores perjudiciales. Por ejemplo, el número de pacientes que un médico ve cada día es un importante motor de la productividad del hospital, pero nadie querría que los médicos se centraran únicamente en ese indicador, y mucho menos que se les evaluara y recompensara solo en función de este. Incluso las mediciones de rendimiento cuantificables —por ejemplo, las ventas de un vendedor o el número de líneas de código escritas de un programador— deben evaluarse en su contexto: no todos los clientes son igual de difíciles de atender, y no todos los proyectos de desarrollo de *software* son idénticos. A la luz de estas realidades, muchas personas no pueden ser evaluadas enteramente sobre la base de mediciones objetivas de rendimiento. De ahí la omnipresencia de las evaluaciones de rendimiento basadas en juicios.[2]

UN CUARTO DE SEÑAL, TRES CUARTOS DE RUIDO

Son miles los artículos de investigación que se han publicado sobre la práctica de la evaluación del rendimiento. La mayoría de los investigadores consideran que estas valoraciones son excesivamente ruidosas.[3] Esta aleccionadora conclusión procede en su mayor parte de estudios basados en revisiones de rendimiento de 360 grados, en las que múltiples calificadores dan su opinión sobre la misma persona que está siendo calificada, en

general en numerosos ámbitos de rendimiento. Cuando se realiza este análisis, el resultado no es el deseado. Los estudios suelen constatar que la verdadera varianza, es decir, la atribuible al rendimiento de la persona, no es más del 20 al 30 por ciento de la varianza total. El resto, entre el 70 y el 80 por ciento de la varianza en las calificaciones, es ruido del sistema.[4]

¿De dónde viene ese ruido? Gracias a numerosos estudios sobre la varianza en las calificaciones del rendimiento laboral, sabemos que todos los componentes del ruido del sistema están presentes.[5]

Estos componentes son bastante fáciles de imaginar en el contexto de una calificación del rendimiento. Consideremos dos calificadoras, Lynn y Mary. Si Lynn es indulgente y Mary es dura, en el sentido de que Lynn da, de media, puntuaciones más altas que Mary a todas las personas evaluadas, entonces tenemos ruido de nivel. Como hemos señalado en nuestro análisis de los jueces, este ruido puede significar que Lynn y Mary se forman ideas realmente diferentes o que las dos calificadoras simplemente utilizan la escala de calificación de forma distinta para expresar la misma impresión.

Ahora bien, si Lynn está evaluando a una persona y resulta que tiene una opinión claramente desfavorable de ella y sus aportaciones, su indulgencia general puede verse contrarrestada por su reacción idiosincrásica (y negativa) hacia ella. Esto es lo que hemos llamado un «patrón estable»: la reacción de un evaluador específico a una persona específica a la que está evaluando. Como el patrón es único para Lynn (y para su enjuiciamiento de esa persona), se trata de una fuente de ruido de patrón.

Por último, puede que Mary haya descubierto que alguien abolló su coche en el aparcamiento de la empresa justo antes de rellenar un formulario de calificación, o puede que Lynn haya recibido una gratificación, sorprendentemente generosa, lo que la puso de un humor extraordinariamente bueno mientras evaluaba el rendimiento. Estos hechos pueden, por supuesto, generar ruido de ocasión.

Diferentes estudios llegan a distintas conclusiones sobre el desglose del ruido del sistema en estos tres componentes (nivel, patrón y ocasión), y sin duda podemos imaginar las razones por las que variará de una organización a otra. Sin embargo, todas las formas de ruido son poco deseables. El mensaje básico que se desprende de esta investigación es sencillo: la mayoría de las valoraciones del rendimiento tienen mucho menos que ver con el rendimiento de la persona valorada de lo que desearíamos. Como se

ha concluido en una revisión, «es probable que la relación entre el rendimiento en el trabajo y las calificaciones de este sea leve o, en el mejor de los casos, incierta».[6]

Además, hay muchas razones por las que las calificaciones en las organizaciones pueden no reflejar la percepción por parte del evaluador del verdadero rendimiento de un empleado.[7] Por ejemplo, es posible que los calificadores no intenten evaluar el rendimiento real, sino que califiquen a las personas «estratégicamente».[8] Entre otras razones, los evaluadores podrían inflar de forma deliberada una calificación para evitar una conversación difícil sobre sus razones, para favorecer a una persona que busca un ascenso largamente esperado o incluso, paradójicamente, para deshacerse de un miembro del equipo de bajo rendimiento que necesita una buena evaluación para que le permitan trasladarse a otra división.

Estos cálculos estratégicos afectan, sin duda, a las calificaciones, pero no son la única fuente de ruido. Lo sabemos gracias a una especie de experimento natural: algunos sistemas de retroalimentación de 360 grados se utilizan solo con fines evolutivos. Con estos sistemas, se dice a los calificadores que la retroalimentación no se utilizará con fines de evaluación. En la medida en que los calificadores creen realmente lo que se les dice, este enfoque les disuade de inflar —o desinflar— las calificaciones. Resulta así que la revisión evolutiva marca una diferencia en la calidad de la retroalimentación, pero el ruido del sistema sigue siendo alto, y continúa representando una varianza mucho mayor que la del rendimiento de la persona calificada. Incluso cuando la retroalimentación es puramente evolutiva, las calificaciones siguen siendo ruidosas.[9]

UN PROBLEMA LARGAMENTE RECONOCIDO, PERO NO RESUELTO

Si los sistemas de calificación del rendimiento dejan tanto que desear, las personas que miden el rendimiento deben tomar nota y mejorarlos. De hecho, en las últimas décadas, las organizaciones han experimentado con innumerables reformas de estos sistemas. Estas han empleado algunas de las estrategias de reducción del ruido que hemos expuesto. En nuestra opinión, se podría hacer mucho más.

Casi todas las organizaciones utilizan para la reducción del ruido la estrategia de la «suma». Las calificaciones sumadas se asocian a menudo a

los sistemas de calificación de 360 grados, que en la década de 1990 se convirtieron en la norma de las grandes empresas. (En 1993, la revista *Human Resources Management* publicó un número especial sobre la retroalimentación de 360 grados.)

Aunque promediar las calificaciones de varios calificadores debería ayudar a reducir el ruido del sistema, hay que precisar que los sistemas de retroalimentación de 360 grados no se inventaron como remedio para ese problema. Su objetivo principal es medir mucho más que lo que ve un jefe. Cuando se pide a empleados y subordinados, y no solo al jefe, que contribuyan a la evaluación de un rendimiento, la naturaleza de lo que se evalúa cambia. La teoría es que este cambio es para mejor, porque los trabajos actuales implican algo más que complacer al jefe. El aumento de la popularidad de la retroalimentación de 360 grados coincidió con la generalización de las organizaciones fluidas, basadas en proyectos.

Algunas pruebas sugieren que la retroalimentación de 360 grados es una herramienta útil en el sentido de que predice el rendimiento objetivamente medible.[10] Por desgracia, el uso de este sistema de retroalimentación ha creado sus propios problemas. A medida que la informatización hacía fácil el incluir más preguntas a los sistemas de retroalimentación, y que la proliferación de múltiples objetivos y limitaciones corporativos añadía aspectos a las descripciones de los trabajos, muchos cuestionarios de retroalimentación se volvieron absurdamente complejos. Abundaban los cuestionarios sobredimensionados (un ejemplo incluye cuarenta y seis valoraciones en once aspectos para cada evaluador y persona evaluada).[11] Se necesitaría un evaluador sobrehumano para recordar y procesar datos precisos y relevantes sobre numerosas personas que están siendo evaluadas en tantos aspectos. Este método excesivamente complicado no solo acaba siendo inútil, sino también funesto. Como hemos visto, el «efecto halo» implica que las dimensiones supuestamente separadas no se tratarán, de hecho, por separado. Una calificación muy positiva o muy negativa en una de las primeras preguntas tenderá a orientar las respuestas a las siguientes en la misma dirección.

Y, lo que es más importante, el desarrollo de los sistemas de 360 grados ha aumentado exponencialmente la cantidad de tiempo que se dedica a proporcionar la retroalimentación. No es infrecuente que se pida a quienes ocupan puestos intermedios que rellenen decenas de cuestionarios sobre sus

colegas de todos los niveles, y, a veces, sobre sus homólogos de otras organizaciones, ya que muchas empresas solicitan ahora retroalimentación de clientes, vendedores y otros socios. Por muy bienintencionado que sea, no cabe esperar que este auge en las exigencias impuestas a los calificadores con limitaciones de tiempo mejore la calidad de la información que suministran. En este caso, puede que la reducción del ruido no merezca la pena (un problema del que trataremos en la sexta parte).

Por último, los sistemas de 360 grados no son inmunes a una enfermedad casi universal de todos los sistemas de medición del rendimiento: la inflación progresiva de las calificaciones. Una gran empresa industrial observó en una ocasión que el 98 por ciento de sus directivos habían sido calificados como «plenamente satisfactorios».[12] Cuando casi todo el mundo recibe la máxima calificación posible, es justo cuestionar el valor de dichas calificaciones.

ELOGIO DE LOS JUICIOS RELATIVOS

Una solución teóricamente eficaz al problema de la inflación de las calificaciones es introducir en ellas cierta estandarización. Una práctica popular que tiene este mismo objetivo es la «clasificación forzada».[13] En un sistema de clasificación forzada, no solo se impide a los calificadores dar a todos la calificación más alta posible, sino que se les obliga a ajustarse a una distribución predeterminada. La clasificación forzada fue defendida por Jack Welch cuando era director ejecutivo (CEO) de General Electric como una forma de detener la inflación en las calificaciones y garantizar la «franqueza» en las evaluaciones de rendimiento. Muchas empresas lo adoptaron y luego lo abandonaron al entender que tenía efectos colaterales poco deseables en la moral y el trabajo en equipo.

Independientemente de sus defectos, las clasificaciones son menos ruidosas que las calificaciones. En el ejemplo de las indemnizaciones punitivas vimos que hay mucho menos ruido en los juicios relativos que en los absolutos, y se ha demostrado que esta relación también se da en las calificaciones del rendimiento.[14]

Para apreciar la razón, obsérvese la figura 17, en la que se muestran dos ejemplos de escalas para evaluar a los empleados. El panel A, en el que se califica a un empleado en una escala absoluta, requiere lo que hemos

llamado una operación de «emparejamiento»: encontrar la puntuación que más se acerque a la impresión que se tenga sobre la «calidad del trabajo» del empleado. En cambio, el panel B requiere que se compare a cada individuo con un grupo de otros en una dimensión específica, la seguridad. Se pide al supervisor que indique el rango (o percentil) de un empleado en una población determinada, utilizando una escala de percentiles. Podemos ver que un supervisor ha colocado a tres empleados en esta escala común.

El método del panel B tiene dos ventajas. En primer lugar, calificar a todos los empleados en una dimensión a la vez (en este ejemplo, la seguridad) ejemplifica una estrategia de reducción del ruido que analizaremos con más detalle en el próximo capítulo: «estructurar» un juicio complejo en varias dimensiones. La estructuración es un intento de limitar el «efecto halo», que suele mantener las valoraciones de un individuo en diferentes dimensiones dentro de un rango pequeño. (Por supuesto, la estructuración solo funciona si la clasificación se realiza en cada dimensión por separado, como en este ejemplo: clasificar a los empleados según un juicio sumado y mal definido sobre la «calidad del trabajo» no reduciría el «efecto halo».)

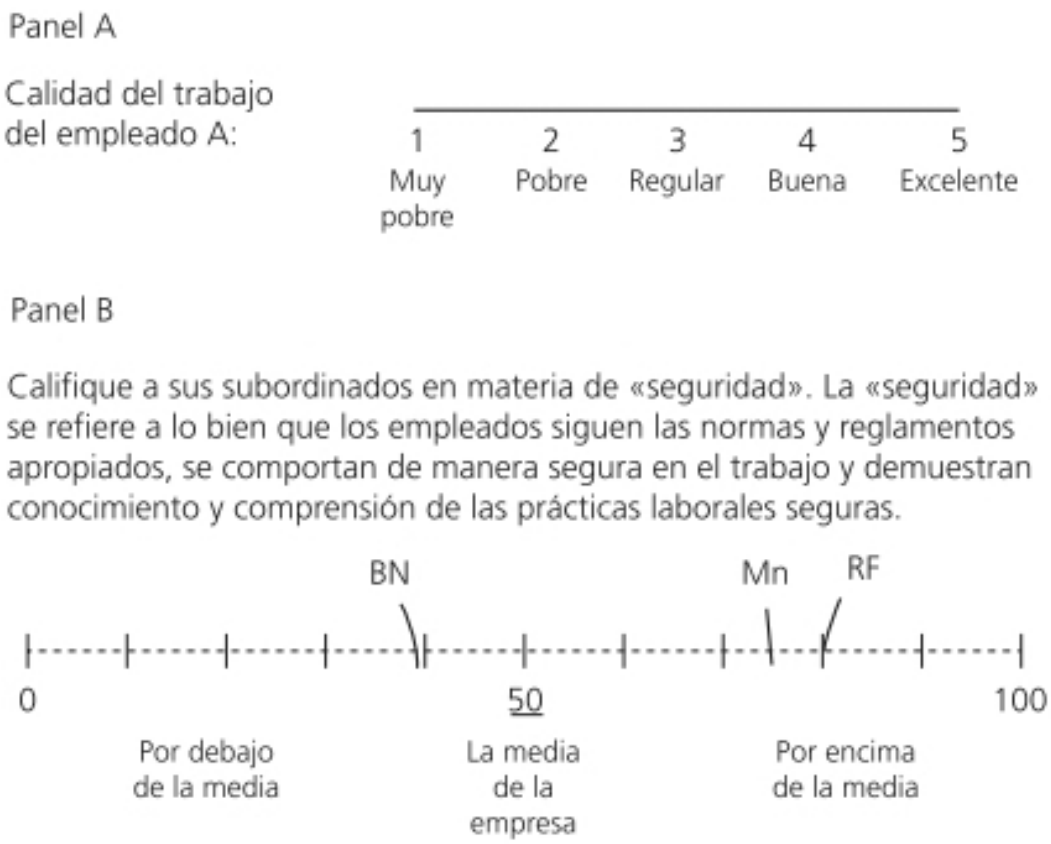

FIGURA 17. Ejemplos de escalas de calificación absoluta y relativa[15]

En segundo lugar, como ya dijimos en el capítulo 15, una clasificación reduce tanto el ruido de patrón como el de nivel. Es menos probable ser incoherente (y crear ruido de patrón) cuando se compara el rendimiento de dos miembros del propio equipo que cuando se da una calificación a cada uno por separado. Y, lo que es más importante, las clasificaciones eliminan mecánicamente el ruido de nivel. Si Lynn y Mary evalúan al mismo grupo de veinte empleados, y Lynn es más indulgente que Mary, sus valoraciones medias serán diferentes, pero no sus clasificaciones medias. Un clasificador indulgente y un clasificador duro utilizan los mismos rangos.

De hecho, la reducción del ruido es el principal objetivo declarado de la clasificación forzada, que garantiza que todos los evaluadores tengan la misma media y la misma distribución de evaluaciones. Las clasificaciones son «forzadas» cuando se impone una distribución de evaluaciones. Por ejemplo, una norma podría establecer que no se puede poner más del 20 por

ciento de las personas evaluadas en la categoría superior y que no se puede poner menos del 15 por ciento en la inferior.

CLASIFICAR, PERO NO FORZAR

En principio, la clasificación forzada aportaría mejoras muy necesarias. Sin embargo, a menudo resulta contraproducente. No pretendemos repasar aquí todos sus posibles efectos no deseados (que suelen estar relacionados con una mala aplicación más que con el principio). Sin embargo, hay dos problemas con los sistemas de clasificación forzada que dan algunas lecciones generales.

La primera es la confusión entre rendimiento absoluto y relativo. Desde luego, es imposible que el 98 por ciento de los gestores de cualquier empresa estén en el 20 por ciento, el 50 por ciento o incluso el 80 por ciento superior de su grupo de colegas. Sin embargo, no es imposible que todos ellos «cumplan las expectativas» si estas se han definido *ex ante* y «en términos absolutos».

Muchos ejecutivos se oponen a la idea de que casi todos los empleados puedan cumplir las expectativas. Si es así, arguyen, las expectativas deben de ser demasiado bajas, tal vez porque reine una cultura de complacencia. Hay que reconocer que esta interpretación puede ser válida, pero también es posible que la mayoría de los empleados cumplan realmente las «altas» expectativas. Esto es exactamente lo que esperaríamos encontrar en una organización de alto rendimiento. Nadie se mofaría de la indulgencia de los procedimientos de gestión del rendimiento que utiliza la NASA si oyera que todos los astronautas de una misión espacial culminada con éxito habían cumplido plenamente las expectativas.

La conclusión es que un sistema que depende de evaluaciones relativas solo es apropiado si una organización se preocupa por el rendimiento relativo. Por ejemplo, las calificaciones relativas pueden tener sentido cuando, independientemente del rendimiento absoluto de las personas, solo un porcentaje fijo de ellas puede ser ascendido (pensemos en los coroneles que son evaluados para su ascenso al generalato). Sin embargo, forzar una clasificación relativa cuando se pretende medir un nivel «absoluto» de rendimiento, como hacen muchas empresas, es ilógico. Y ordenar que un porcentaje determinado de empleados sea calificado como que no cumple

las expectativas (absolutas) no solo es cruel; es absurdo. Sería una insensatez decir que el 10 por ciento de una unidad de élite del ejército ha de ser calificada de «insatisfactoria».

El segundo problema es que se supone que la distribución forzada de las calificaciones refleja la distribución de los verdaderos rendimientos subyacentes (por regla general, algo próximo a una distribución normal). Sin embargo, aunque se conozca la distribución de los rendimientos en la población que se califica, es posible que no se reproduzca la misma distribución en un grupo más pequeño, como los evaluados por un solo evaluador. Si se eligen al azar diez personas de una población de varios miles, no hay garantía de que con exactitud dos de ellas pertenezcan al 20 por ciento más capaz de la población general. («No hay garantía» es quedarse corto: la probabilidad de que esto ocurra es del 30 por ciento.) En la práctica, el problema es aún más serio, porque la composición de los equipos no es aleatoria. Algunas unidades pueden estar formadas casi en su totalidad por empleados de alto rendimiento, y otras por empleados mediocres.

La clasificación forzada en un entorno de este tipo es inevitablemente una fuente de error e injusticia. Supongamos que el equipo asignado a un evaluador está compuesto por cinco personas cuyos rendimientos son indistinguibles. Forzar una distribución diferenciada de calificaciones sobre esta realidad indiferenciada no reduce el error, sino que lo aumenta.

Los críticos de la clasificación forzada han centrado a menudo sus ataques en el principio de la clasificación, al que denuncian como brutal, inhumano y, a la postre, contraproducente. Se acepten o no estos argumentos, el error fatal de la clasificación forzada no es la «clasificación», sino el «forzado». Siempre que se fuerzan los juicios en una escala inapropiada, sea porque se utiliza una escala relativa para medir un rendimiento absoluto o porque se obliga a los juzgadores a distinguir lo indistinguible, la elección de la escala añade, de forma automática, ruido.

¿QUÉ ES LO SIGUIENTE?

A la luz de todos los esfuerzos que las organizaciones han hecho para mejorar la medición del rendimiento, decir que los resultados han sido decepcionantes es quedarse corto. Como resultado de esos esfuerzos, el

coste de las evaluaciones de rendimiento se disparó. En 2015, Deloitte calculó que se dedicaban dos millones de horas al año a evaluar a sesenta y cinco mil personas.[16] Las revisiones del rendimiento siguen siendo uno de los rituales más temidos de las organizaciones, odiado casi tanto por quienes tienen que realizarlas como por quienes son objeto de ellas. Un estudio reveló que un pasmoso 90 por ciento de los directivos, empleados y responsables de recursos humanos creen que sus procesos de gestión del rendimiento no consiguen los resultados esperados.[17] La investigación ha confirmado lo que la mayoría de los directivos han experimentado. Aunque la retroalimentación sobre el rendimiento cuando se asocia a un plan evolutivo para el empleado puede producir mejoras, las evaluaciones del rendimiento más comúnmente practicadas desmotivan con la misma frecuencia que motivan. Como concluía un artículo de revisión, «pese a lo que se ha intentado hacer durante décadas para mejorar los procesos [de gestión del rendimiento], estos siguen generando información inexacta y no hacen casi nada para impulsar el rendimiento».[18]

En su desesperación, un pequeño pero creciente número de empresas está considerando la opción radical de eliminar por completo los sistemas de evaluación. Los defensores de esta «revolución en la gestión del rendimiento», entre los que se encuentran muchas empresas tecnológicas, algunas organizaciones de servicios profesionales y un puñado de empresas de sectores tradicionales, pretenden centrarse en la retroalimentación orientada a la evolución y al futuro, en lugar de en la evaluación retrospectiva.[19] Algunos incluso han hecho que sus evaluaciones no tengan números, lo que significa que abandonan las calificaciones de rendimiento tradicionales.

En cuanto a las empresas que no renuncian a las calificaciones del rendimiento (y son la inmensa mayoría), ¿qué se puede hacer para mejorarlas? Una estrategia de reducción del ruido tiene que ver, de nuevo, con la elección de la escala adecuada. El objetivo es garantizar un «marco de referencia común». La investigación sugiere que una combinación de formatos de calificación mejorados y la formación de los evaluadores puede ayudar a lograr una mayor coherencia entre los evaluadores en su uso de la escala.

Como mínimo, las escalas de valoración del rendimiento deben estar ancladas en descriptores lo suficientemente específicos como para ser

interpretados de forma coherente. Muchas organizaciones utilizan «escalas de valoración ancladas en el comportamiento», en las que cada grado de la escala corresponde a una descripción de comportamientos específicos. El panel izquierdo de la figura 18 ofrece un ejemplo.

Sin embargo, las pruebas indican que las escalas de valoración ancladas en el comportamiento no bastan para eliminar el ruido.[20] Se ha demostrado que un paso más, la «preparación en el marco de referencia», ayuda a garantizar la coherencia entre los evaluadores. En este paso, se prepara a los evaluadores para que reconozcan los diferentes aspectos del rendimiento. Practican la evaluación del rendimiento mediante viñetas grabadas en vídeo y luego aprenden a comparar sus evaluaciones con las «verdaderas» proporcionadas por expertos.[21] Las viñetas de rendimiento actúan como casos de referencia; cada viñeta define un punto de anclaje en la escala de rendimiento, que se convierte en una «escala de casos» como la que se muestra en el panel derecho de la figura 18.

Con una escala de casos, cada evaluación de un nuevo individuo es una comparación con los casos de anclaje. Se convierte así en un juicio relativo. Dado que los juicios comparativos son menos susceptibles al ruido que las calificaciones, las escalas de casos son más fiables que las escalas que utilizan números, adjetivos o descripciones del comportamiento.

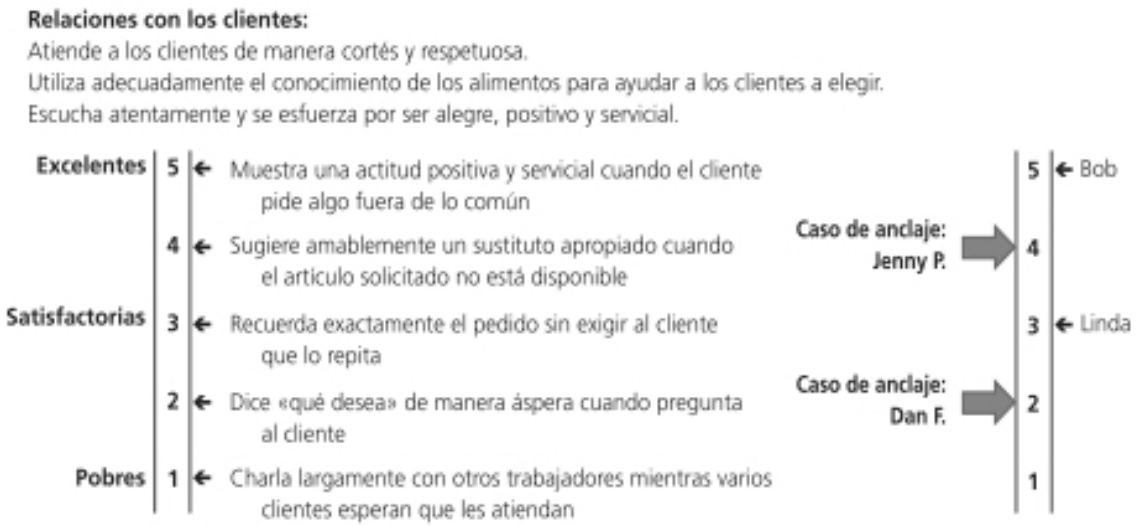

FIGURA 18. Ejemplo de escala de calificación anclada en el comportamiento (izquierda) y escala de casos (derecha)[22]

La preparación en el marco de referencia se conoce desde hace décadas y permite calificaciones que, como se ha demostrado, son menos ruidosas y más exactas. Sin embargo, ha ganado poco terreno. Es fácil adivinar por qué. La preparación en el marco de referencia, las escalas de casos y otras herramientas que persiguen los mismos objetivos son complejas y requieren mucho tiempo. Para que sean valiosas, suelen necesitar adaptarse a la empresa e incluso a la unidad que realiza las evaluaciones, y deben actualizarse con frecuencia a medida que evolucionan los requisitos del puesto. Estas herramientas requieren que la empresa las sume a su ya cuantiosa inversión en sus sistemas de gestión del rendimiento. Los usos actuales van en la dirección opuesta. (En la sexta parte, tendremos más que decir sobre los costes de la reducción del ruido.)

Además, cualquier organización que controle el ruido atribuible a los calificadores también reduce su capacidad para influir en las calificaciones en pos de sus propios objetivos. Exigir a los directivos que reciban una formación añadida para sus calificaciones, que inviertan más esfuerzo en el proceso de calificación y que renuncien a parte del control que tienen sobre los resultados es algo que inevitablemente genera una gran resistencia. Resulta revelador que la mayoría de los estudios sobre la preparación de las evaluaciones en un marco de referencia se hayan realizado hasta ahora con estudiantes, no con directivos en ejercicio.[23]

El amplio tema de la evaluación del rendimiento plantea muchas cuestiones, tanto prácticas como filosóficas. Por ejemplo, hay quienes se preguntan hasta qué punto la noción de «rendimiento individual» tiene sentido en las organizaciones actuales, en las que los resultados dependen a menudo de cómo interactúan las personas entre sí. Si creemos que la noción tiene sentido de verdad, debemos preguntarnos cómo se distribuyen los niveles de rendimiento individual entre las personas de una determinada organización; por ejemplo, si el rendimiento tiene una distribución normal o si existen «talentos estrella» que hacen una contribución enorme y hasta desproporcionada.[24] Y si el objetivo es sacar lo mejor de las personas, cabe preguntarse razonablemente si medir el rendimiento individual y utilizar esa medición para motivar a la gente mediante el temor y la codicia es el mejor método (o incluso si es un método eficaz).

Si está diseñando o revisando un sistema de gestión del rendimiento, es necesario responder a estas preguntas y a otras muchas más. No aspiramos

aquí a darles respuesta, sino a hacer una sugerencia más discreta: decir a quienes miden el rendimiento que es probable que sus calificaciones estén afectadas por ruido del sistema y, por esa razón, puedan ser en esencia inútiles y muy posiblemente contraproducentes. Reducir este ruido es un reto que no puede resolverse con simples arreglos tecnológicos. Requiere una reflexión clara sobre los juicios que se espera que hagan los evaluadores. Lo más probable es que se puedan mejorar las valoraciones clarificando la escala de calificaciones y formando a los evaluadores para que la utilicen de forma coherente. Esta estrategia de reducción del ruido es aplicable en muchos otros campos.

SOBRE DEFINIR LA ESCALA

- «Dedicamos mucho tiempo a nuestras calificaciones del rendimiento y, sin embargo, los resultados son una cuarta parte de rendimiento y tres cuartas partes de ruido del sistema.»
- «Probamos la retroalimentación de 360 grados y forzamos la clasificación para abordar este problema, pero puede que hayamos empeorado las cosas.»
- «Si hay tanto ruido de nivel, es porque diferentes evaluadores tienen ideas completamente distintas de lo que significa “bueno” o “excelente”. Solo estarán de acuerdo si les damos casos concretos como anclas en la escala de valoración.»

24
Estructura en la contratación

Quien una vez obtuvo un empleo del tipo que fuere, las palabras «entrevista de trabajo» pueden evocarle algunos recuerdos vívidos y estresantes. Las entrevistas de trabajo en las que un candidato se reúne con un futuro supervisor o un profesional de recursos humanos son un rito de paso necesario para entrar en muchas organizaciones.

En la mayoría de los casos, las entrevistas siguen una rutina muy ensayada. Tras intercambiar unas palabras de cortesía, los entrevistadores piden a los candidatos que describan su experiencia o que se explayen sobre aspectos concretos de esta. Les hacen preguntas sobre sus logros y retos, sobre sus motivaciones para acceder al puesto de trabajo o sobre sus ideas para mejorar la empresa. A menudo, los entrevistadores piden a los candidatos que describan su personalidad y que expliquen por qué encajarían bien en el puesto o en la cultura de la empresa. A veces se habla de las aficiones y de los intereses. Hacia el final, el candidato suele hacer algunas preguntas, que se evalúan debidamente en cuanto a su pertinencia y perspicacia.

Si el entrevistador está en condiciones de contratar empleados, sus métodos de selección probablemente incluyan alguna versión de este ritual. Como señaló un psicólogo de las organizaciones, «es raro, incluso impensable, que alguien sea contratado sin algún tipo de entrevista».[1] Y casi todos los profesionales confían en cierta medida en sus juicios intuitivos a la hora de decidirse a contratar empleados después de estas entrevistas.[2]

La omnipresencia de la entrevista de trabajo refleja una creencia muy arraigada en el valor del juicio a la hora de elegir a las personas con las que se va a trabajar. Y, como tarea de juicio, la selección de personal tiene una gran ventaja: al ser tan omnipresente y tan importante, los psicólogos de las organizaciones la han estudiado con gran detalle. El primer número del *Journal of Applied Psychology*, publicado en 1917, identificaba la contratación como el «problema supremo [...] porque las capacidades

humanas son, después de todo, los principales recursos nacionales».[3] Un siglo después, sabemos mucho sobre la eficacia de las distintas técnicas de selección (incluidas las entrevistas estándar). Ninguna tarea de juicio compleja ha sido objeto de tanta investigación de campo. Esto hace de ella un caso de prueba perfecto que ofrece lecciones que pueden extrapolarse a muchos juicios que implican una elección entre varias opciones.

LOS PELIGROS DE LAS ENTREVISTAS

A quien no esté familiarizado con la investigación sobre las entrevistas de trabajo, lo que sigue puede sorprenderle. Si el objetivo de las entrevistas es determinar qué candidatos tendrán éxito en un puesto de trabajo y qué otros fracasarán, las entrevistas estándar (también llamadas «entrevistas no estructuradas», para distinguirlas de las «estructuradas», a las que nos referiremos en breve) no son muy informativas. Para decirlo más claramente, a menudo son inútiles.

Para llegar a esta conclusión, innumerables estudios estimaron la correlación entre la calificación que un evaluador da a un candidato después de una entrevista y el éxito del candidato en el trabajo. Si la correlación entre la calificación de la entrevista y el éxito es alta, se puede suponer que las entrevistas —o cualquier otra técnica de contratación para la que se calcule la correlación de la misma manera— son un buen predictor del rendimiento de los candidatos.

Es necesario hacer aquí una advertencia. La definición de «éxito» no es un problema en absoluto trivial. El rendimiento se evalúa por regla general sobre la base de las calificaciones de los supervisores. A veces, la medición es la duración del empleo. Por supuesto, estas mediciones plantean dudas, sobre todo teniendo en cuenta la cuestionable validez de las calificaciones del rendimiento a que nos hemos referido en el capítulo anterior. Sin embargo, a efectos de evaluar la calidad de los juicios de un empleador a la hora de seleccionar a sus empleados, parece razonable utilizar los juicios que el mismo empleador hace al evaluar a los empleados así contratados. Cualquier análisis de la calidad de las decisiones de contratación debe partir de este supuesto.

¿Qué conclusiones se extraen de estos análisis? En el capítulo 11 mencionamos una correlación entre las evaluaciones típicas de las

entrevistas y las evaluaciones del desempeño del trabajo de .28. Otros estudios informan de correlaciones que oscilan entre .20 y .33.[4] Como hemos visto, se trata de una correlación muy buena según los estándares de las ciencias sociales, pero no muy buena para basar en ella las decisiones. Utilizando el porcentaje de concordancia (PC) que introdujimos en la tercera parte del libro, podemos calcular una probabilidad: dados los niveles de correlación anteriores, si todo lo que se sabe sobre dos candidatos es que uno parecía mejor que el otro en las entrevistas, las posibilidades de que este candidato tenga en efecto un mejor rendimiento son de entre el 56 y el 61 por ciento. Algo mejor que lanzar una moneda al aire, sin duda, pero no es una forma segura de tomar decisiones importantes.

Es cierto que las entrevistas sirven para otros propósitos, además de emitir un juicio sobre un candidato. En particular, ofrecen la oportunidad de dejar la empresa en manos de candidatos prometedores y de empezar a establecer relaciones con futuros colegas. Sin embargo, desde la perspectiva de una organización que invierte tiempo y esfuerzo en la selección de talentos, el objetivo principal de las entrevistas es claramente la selección. Y, en esa tarea, no tienen precisamente un éxito rotundo.

EL RUIDO EN LAS ENTREVISTAS

Podemos ver fácilmente por qué las entrevistas tradicionales producen errores en su predicción del rendimiento laboral. Parte de este error tiene que ver con lo que hemos denominado «ignorancia objetiva» (véase el capítulo 11).[5] El rendimiento en el trabajo depende de muchas cosas, como la rapidez con que la persona contratada se adapte a su nuevo puesto o cómo afecten a su trabajo algunos acontecimientos de su vida. Gran parte de esto es impredecible en el momento de la contratación. Esta incertidumbre limita la validez predictiva de las entrevistas y, no menos, de cualquier otra técnica de selección de personal.

Las entrevistas son también un campo minado de sesgos psicológicos. En los últimos años se ha observado que los entrevistadores tienden, a menudo sin quererlo, a preferir a los candidatos que son culturalmente similares a ellos[6] o a aquellos con los que tienen algo en común, como el sexo, la raza y la formación. Muchas empresas reconocen ahora los riesgos que acarrean los sesgos, e intentan enfrentarse a ellos mediante la preparación

específica de los profesionales de la contratación y otros empleados. También se conocen otros sesgos desde hace décadas. Por ejemplo, el aspecto físico desempeña un papel importante en la evaluación de los candidatos, incluso para puestos en los que debería importar poco o nada. Estos sesgos son compartidos por todos o la mayoría de los contratadores, y, cuando se aplican a un candidato determinado, tienden a producir un error compartido (un sesgo negativo o positivo en la evaluación del candidato).

A nadie sorprenderá saber que también hay ruido: diferentes entrevistadores reaccionan de forma diferente ante el mismo candidato y llegan a diferentes conclusiones. Las mediciones de correlación entre las valoraciones que dos entrevistadores producen tras entrevistar al mismo candidato varían entre .37 y .44 (PC = 6265 %).[7] Una de las razones es que el candidato puede no comportarse exactamente igual con distintos entrevistadores. Sin embargo, incluso en las entrevistas en grupo, en las que varios entrevistadores observan el mismo comportamiento del entrevistado, la correlación entre sus valoraciones dista mucho de ser completa. Un metaanálisis estima una correlación de .74 (PC = 76 %). Esto significa que uno y otro entrevistador, después de ver a los «mismos» dos candidatos en la «misma» entrevista en grupo, seguirán discrepando una cuarta parte del tiempo sobre cuál de los dos candidatos es mejor.

Esta variabilidad es en gran parte producto del ruido de patrón, es decir, de la diferencia en las reacciones idiosincrásicas de los entrevistadores ante un entrevistado. La mayoría de las organizaciones prevén esta variabilidad, y por eso exigen que varios entrevistadores se reúnan con el mismo candidato, y que los resultados se sumen de algún modo. (Normalmente, la opinión sumada se forma por medio de una discusión en la que hay que alcanzar algún tipo de consenso; un procedimiento que crea sus propios problemas, como ya hemos señalado.)

Un descubrimiento más sorprendente es la presencia de mucho ruido de ocasión en las entrevistas. Por ejemplo, hay pruebas sólidas de que las recomendaciones de contratación están vinculadas con las impresiones que se crean en la fase de relaciones informales de una entrevista, esos dos o tres primeros minutos en los que se charla amistosamente para que el candidato se sienta cómodo. La primera impresión resulta ser importante (y mucho).[8]

Tal vez se piense que juzgar por las primeras impresiones no es problemático. Al menos parte de lo que aprendemos de las primeras impresiones es significativo. Todos sabemos que aprendemos algo en los primeros segundos de interacción con una persona que desconocíamos. Es lógico que esto sea particularmente cierto en el caso de los entrevistadores cualificados. Pero los primeros segundos de una entrevista reflejan exactamente el tipo de cualidades superficiales que se asocian a las primeras impresiones: las primeras percepciones se basan sobre todo en la extraversión y las habilidades verbales del candidato. Incluso la calidad de un apretón de manos es un predictor significativo de las recomendaciones de contratación.[9] Puede que a todos nos guste un apretón de manos firme, pero pocos contratadores harían conscientemente de él un criterio clave de contratación.

LA PSICOLOGÍA DE LOS ENTREVISTADORES

¿Por qué las primeras impresiones acaban determinando el resultado de una entrevista mucho más larga? Una de las razones es que, en una entrevista tradicional, los entrevistadores tienen libertad para dirigir la entrevista en la dirección que consideren oportuna. Es probable que hagan preguntas que confirmen una impresión inicial. Si un candidato parece tímido y reservado, por ejemplo, el entrevistador puede querer hacer preguntas difíciles sobre las experiencias pasadas del candidato en el trabajo en equipo, pero tal vez se olvide de realizar las mismas preguntas a alguien que parece jovial y sociable. Las pruebas recogidas sobre estos dos candidatos no serán las mismas. Un estudio que siguió el comportamiento de los entrevistadores que habían tenido una impresión inicial positiva o negativa a partir de los currículos y las puntuaciones de los test descubrió que las impresiones iniciales tienen un profundo efecto en el desarrollo de la entrevista. Los entrevistadores con una primera impresión positiva, por ejemplo, hacen menos preguntas y tienden a «vender» la empresa al candidato.[10]

El poder de la primera impresión no es el único aspecto problemático de las entrevistas. Otro es que, como entrevistadores, queremos que el candidato que tenemos delante «tenga sentido» (una manifestación de nuestra excesiva tendencia, analizada en el capítulo 13, a buscar y encontrar la coherencia). En un sorprendente experimento, los investigadores

asignaron a unos estudiantes el papel de entrevistador o entrevistado, y dijeron a unos y a otros que la entrevista debía consistir solo en preguntas cerradas cuya respuesta fuese un sí o un no. A continuación, pidieron a algunos de los entrevistados que respondieran «aleatoriamente» a esas preguntas.[11] (La primera letra de las preguntas, tal y como estaban formuladas, determinaba si debían responder sí o no.) Como comentaron irónicamente los investigadores, «algunos de los entrevistados estaban inicialmente preocupados por que la entrevista aleatoria se rompiera y se revelara como un sinsentido. No ocurrió así, y las entrevistas continuaron». La sorpresa fue que «ni un solo entrevistador» se dio cuenta de que los candidatos estaban dando respuestas al azar. Peor aún, cuando se les pidió que estimaran si eran «capaces de inferir muchas cosas sobre esa persona dado el tiempo que pasaron juntos», los entrevistadores en esta condición «aleatoria» estuvieron de acuerdo con la misma probabilidad que los que habían conocido a los candidatos que respondían verdaderamente. Así es nuestra capacidad de crear coherencia. Igual que a menudo podemos encontrar un patrón imaginario en datos aleatorios o imaginar una forma en los contornos de una nube, somos capaces de encontrar una lógica en respuestas totalmente carentes de sentido.

Para un ejemplo menos extremo, consideremos el siguiente caso. Uno de los autores de este libro tuvo que entrevistar a un candidato que fue, en su puesto anterior, director financiero de una empresa mediana. Advirtió que el candidato había dejado ese puesto al cabo de unos meses y le preguntó por qué. El candidato le explicó que el motivo era un «desacuerdo estratégico con el director ejecutivo (CEO)». Un colega también entrevistó al candidato, le hizo la misma pregunta y obtuvo la misma respuesta. Sin embargo, en el informe que siguió a estas entrevistas, los dos entrevistadores tuvieron opiniones radicalmente diferentes. Uno de ellos, que había hecho una evaluación positiva del candidato, consideró que la decisión que este tomó de abandonar la empresa era una muestra de integridad y valentía. El otro, que había tenido una primera impresión negativa, interpretó el mismo hecho como un signo de inflexibilidad, tal vez incluso de inmadurez. Esta historia ilustra que, por mucho que queramos creer que nuestro juicio sobre un candidato se basa en hechos, en nuestra interpretación de los hechos se traslucen nuestras actitudes previas.

Las limitaciones de las entrevistas tradicionales arrojan serias dudas sobre nuestra capacidad para sacar conclusiones de alguna significación. Sin embargo, las impresiones que se tienen en una entrevista son vívidas, y el entrevistador suele confiar en ellas. Cuando se combinan las conclusiones a que se llega en una entrevista con otras pistas sobre el candidato, tendemos a dar demasiada importancia a la entrevista y muy poca a otros datos que pueden ser más predictivos, como las puntuaciones en ciertos test.

Una historia puede ayudar a concretar esta observación. A menudo se pide a profesores que son entrevistados para acceder a un puesto en una facultad que impartan clases ante un plantel de otros profesores para garantizar que su capacidad docente está a la altura de la institución. Esta es, por supuesto, una situación de mayor riesgo que una clase ordinaria. En una ocasión, uno de nosotros fue testigo de cómo un candidato causaba una mala impresión en este ejercicio, y estaba claro que esta se debía al estrés de la propia situación: el currículo del candidato mencionaba evaluaciones docentes sobresalientes y varios premios a la excelencia docente. Sin embargo, la vívida impresión que producía su fracaso en una situación muy artificial pesó más en la decisión final que los datos abstractos sobre su excelente capacidad docente en el pasado.

Una última puntualización: cuando las entrevistas no son la única fuente de información sobre los candidatos —por ejemplo, cuando también hay test, referencias u otros datos—, estos diversos *inputs* deben combinarse en un juicio global. La cuestión que aquí se plantea ya la conocemos: ¿deben combinarse esos datos utilizando el juicio (una suma clínica) o una fórmula (una suma mecánica)? Como vimos en el capítulo 9, el enfoque mecánico es superior tanto en general como en el caso específico de la predicción del rendimiento laboral. Lamentablemente, las encuestas dicen que la inmensa mayoría de los profesionales de recursos humanos están a favor de la suma clínica. Esta práctica añade otra fuente de ruido a un proceso ya de por sí ruidoso.[12]

LA MEJORA DE LA SELECCIÓN DE PERSONAL MEDIANTE LA ESTRUCTURACIÓN

Si las entrevistas tradicionales y las decisiones de contratación basadas en el juicio tienen una validez predictiva limitada, ¿qué podemos hacer al respecto? Afortunadamente, la investigación también ha ofrecido algunos

consejos sobre cómo mejorar la selección de personal, y algunas empresas les prestan atención.

Un ejemplo de empresa que ha mejorado sus prácticas de selección de personal y ha informado de sus resultados es Google. Laszlo Bock, su antiguo vicepresidente senior de People Operations, lo cuenta en su libro *Work Rules*! A pesar de centrarse en la contratación de talentos del más alto nivel y de dedicar considerables recursos a la búsqueda de las personas adecuadas, Google tenía problemas. Una auditoría de la validez predictiva de sus entrevistas de contratación encontró «una relación cero [...], un completo desorden aleatorio».[13] Los cambios que Google puso en práctica para afrontar esta situación reflejan principios que han surgido de décadas de investigación. También ilustran las estrategias de higiene de las decisiones.

Una de estas estrategias ya debería sernos familiar: la suma. Su uso en este contexto no es una sorpresa. Casi todas las empresas suman las opiniones de varios entrevistadores sobre el mismo candidato. Para no quedarse atrás, Google a veces hacía que los candidatos sufrieran veinticinco entrevistas. Una de las conclusiones de la revisión de Bock fue reducir ese número a cuatro, ya que descubrió que las entrevistas nuevas no añadían casi ninguna validez predictiva a lo conseguido con las cuatro primeras. Sin embargo, con el fin de garantizar este nivel de validez, Google aplica estrictamente una norma que no todas las empresas observan: la empresa se asegura de que los entrevistadores califiquen al candidato por separado «antes» de que estos se comuniquen entre sí. Una vez más: la suma funciona, pero solo si los juicios son independientes.

Google adoptó también una estrategia de higiene de las decisiones que aún no hemos descrito en detalle: la «estructuración de juicios complejos». El término «estructura» puede significar muchas cosas. Tal y como aquí utilizamos el término, a un juicio complejo estructurado lo definen tres principios: descomposición, independencia y juicio holístico retardado.

El primer principio, el de la «descomposición», desglosa la decisión en componentes, o «evaluaciones mediadoras». Este paso tiene el mismo propósito que la identificación de los subjuicios en una directriz: centra a los juzgadores en las pistas importantes. La descomposición actúa como una hoja de ruta para especificar qué datos se necesitan, y filtra la información irrelevante.

En el caso de Google, hay cuatro evaluaciones mediadoras en la descomposición: la capacidad cognitiva general, el liderazgo, el ajuste cultural (llamado «googleyness») y los conocimientos relacionados con el papel. (Algunas de estas evaluaciones se desglosan luego en componentes más pequeños.) En ellas puede observarse que la buena presencia de un candidato, su facilidad de palabra, sus interesantes aficiones y cualquier otro aspecto, positivo o negativo, en el que un contratador pueda fijarse en una entrevista no estructurada no están en la lista.

Crear este tipo de estructura para una tarea de contratación puede parecer de simple sentido común. De hecho, si se va a contratar a un contable de nivel básico o a un auxiliar administrativo, existen descripciones estándar de puestos que especifican las competencias necesarias. Sin embargo, como saben los contratadores profesionales, la definición de las evaluaciones clave se hace difícil para los puestos poco habituales o de alto nivel, y es frecuente que este paso de la definición se omita. Un destacado cazatalentos señala que definir las competencias requeridas de manera suficientemente específica es una tarea difícil y que a menudo se pasa por alto. Destaca la importancia para los responsables de la toma de decisiones de «invertir en la definición del problema»: dedicar el tiempo necesario por adelantado, antes de conocer a ningún candidato, para acordar una descripción clara y detallada del puesto.[14] El problema que aquí se plantea es que muchos entrevistadores utilizan descripciones hinchadas de puestos de trabajo establecidas por consenso y compromiso. Las descripciones son vagas listas de todas las características que debe reunir el candidato ideal, y no ofrecen ningún modo de calibrar esas características o de buscar compensaciones entre ellas.

El segundo principio del juicio estructurado, la «independencia», requiere que la información sobre cada evaluación se recoja de forma aislada. No basta con enumerar los componentes de la descripción del puesto: la mayoría de los contratadores que realizan entrevistas tradicionales también conocen las cuatro o cinco cosas que buscan en un candidato. El problema es que, al realizar la entrevista, no evalúan estos elementos por separado. Cada evaluación influye en las demás, lo que hace que cada evaluación sea muy ruidosa.

Para superar este problema, Google organizó formas de hacer evaluaciones basadas en hechos e independientes entre sí. Tal vez su

movimiento más visible fue la introducción de «entrevistas estructuradas sobre el comportamiento».[15] La tarea de los entrevistadores en este tipo de entrevistas no es decidir si les gusta un candidato en general, sino recoger datos sobre cada estimación en la estructura de evaluación y asignar una puntuación al candidato en cada evaluación. Para ello, los entrevistadores deben hacer preguntas definidas previamente sobre el comportamiento del candidato en situaciones anteriores. También han de registrar las respuestas y puntuarlas según una escala de valoración predeterminada, utilizando una rúbrica unificada. La rúbrica da ejemplos de cómo son las respuestas medias, buenas o excelentes a cada pregunta. Esta escala compartida (un ejemplo de las escalas de calificación ancladas en el comportamiento que presentamos en el capítulo anterior) ayuda a reducir el ruido en los juicios.

Si este enfoque se antoja diferente a una entrevista tradicional, es porque en efecto lo es. De hecho, puede parecerse más a un examen o a un interrogatorio que a un encuentro en un marco empresarial, y hay algunas pruebas de que, tanto a los entrevistados como a los entrevistadores, les disgustan las entrevistas estructuradas (o, al menos, prefieren las no estructuradas). Se sigue debatiendo sobre qué debe incluir exactamente una entrevista para que se la califique de estructurada.[16] Aun así, una de las conclusiones más consistentes que se desprenden de las publicaciones sobre las entrevistas es que las entrevistas estructuradas son mucho más predictivas del rendimiento futuro que las tradicionales no estructuradas. Las correlaciones con el rendimiento laboral oscilan entre .44 y .57. Utilizando nuestra métrica de PC, las posibilidades de elegir al mejor candidato con una entrevista estructurada se sitúan entre el 65 y el 69 por ciento, lo que supone una notable mejora con respecto a las posibilidades del 56 al 61 por ciento de una entrevista no estructurada.[17]

Google utiliza otros datos como *inputs* en algunos de los aspectos que le interesan. Para comprobar los conocimientos relacionados con el trabajo, se basa en parte en «test de muestras de trabajo», como pedir a un candidato a un puesto de programación que escriba algún código. La investigación ha demostrado que los test de muestras de trabajo están entre los mejores predictores del rendimiento en el trabajo.[18] Google también utiliza «referencias de puerta trasera», que no son proporcionadas por alguien que

el candidato haya nombrado, sino por empleados de Google con los que el candidato se haya cruzado.

El tercer principio del juicio estructurado, el «juicio holístico retardado», puede resumirse en una simple prescripción: no excluir la intuición, sino retrasarla. En Google, la recomendación final de contratación se realiza de forma colegiada por un comité de contratación que revisa un expediente completo con todas las calificaciones que los candidatos han obtenido en cada evaluación de cada entrevista y otra información relevante en apoyo de esas evaluaciones. Sobre la base de esta información, el comité decide si se hace una oferta.

A pesar de la famosa cultura basada en los datos que distingue a esta empresa, y a pesar de todas las pruebas de que una combinación mecánica de datos supera a una clínica, la decisión final de contratación «no» es mecánica. Sigue siendo un juicio, un juicio en el que el comité tiene en cuenta todas las pruebas y las sopesa de forma holística, iniciando un debate sobre la pregunta: «¿Tendrá éxito esta persona en Google?». La decisión no es algo meramente calculado.

En el próximo capítulo explicaremos por qué creemos que este enfoque para tomar la decisión final es sensato. Sin embargo, nótese que, si bien las decisiones finales de contratación de Google no son mecánicas, se basan en la puntuación media asignada por los cuatro entrevistadores. También cuentan las pruebas subyacentes. En otras palabras, Google admite el juicio y la intuición en el proceso de la toma de decisiones solo después de haber reunido y analizado todas las pruebas. De este modo, mantiene a raya la tendencia de cada entrevistador (y miembro del comité de contratación) a formarse ideas rápidas e intuitivas y a precipitarse en su juicio.

Los tres principios —una vez más, la descomposición, la evaluación independiente en cada aspecto y el juicio holístico retardado— no proporcionan forzosamente una plantilla para todas las organizaciones que intentan mejorar sus procesos de selección. Sin embargo, los principios coinciden en líneas generales con las recomendaciones que los psicólogos de las organizaciones han formulado a lo largo de los años. De hecho, los principios guardan cierto parecido con el método de selección que uno de nosotros (Kahneman) puso en práctica en el ejército israelí ya en 1956, y que describe en *Pensar rápido, pensar despacio*.[19] Este proceso, al igual que el que puso en marcha Google, formalizó una estructura de evaluación

(la lista de aspectos de personalidad y competencia que había que evaluar). Los entrevistadores debían obtener pruebas objetivas relacionadas con cada uno de los aspectos y puntuarlos antes de pasar al siguiente. Y permitía a los contratadores utilizar su juicio y su intuición para alcanzar una decisión final, pero solo después de la evaluación estructurada.

Hay pruebas abrumadoras de la superioridad de los procesos de juicio estructurado (incluidas las entrevistas estructuradas) en la contratación. Existen consejos prácticos para guiar a los ejecutivos que quieran adoptarlos.[20] Como ilustra el ejemplo de Google, y como han señalado otros investigadores, los métodos de juicio estructurado también son menos costosos, porque pocas cosas son tan costosas como el tiempo de las entrevistas presenciales.

Sin embargo, la mayoría de los ejecutivos siguen convencidos del valor insustituible de los métodos informales basados en entrevistas. Sorprendentemente, lo mismo ocurre con muchos candidatos que creen que solo una entrevista cara a cara les permitirá mostrar a un posible empleador su verdadera valía. Los investigadores han llamado a esto «la persistencia de una ilusión».[21] Una cosa está clara: los contratadores y los candidatos subestiman seriamente el ruido en los juicios de contratación.

SOBRE LA ESTRUCTURA EN LA CONTRATACIÓN

- «En las entrevistas tradicionales e informales, a menudo tenemos una sensación irresistible e intuitiva de entender al candidato y saber si la persona dará la talla. Debemos aprender a desconfiar de esa sensación.»
- «Las entrevistas tradicionales no solo son peligrosas por los sesgos, sino también por el ruido.»
- «Debemos estructurar nuestras entrevistas y, más ampliamente, nuestros procesos de selección. Empecemos por definir de forma mucho más clara y específica lo que buscamos en los candidatos, y asegurémonos de evaluar a los candidatos de forma independiente en cada uno de estos aspectos.»

25
El protocolo de evaluaciones mediadoras

Hace algún tiempo, dos de nosotros (Kahneman y Sibony), junto con nuestro amigo Dan Lovallo, describimos un método de toma de decisiones en las organizaciones. Llamamos al método, que fue diseñado con la mitigación del ruido como objetivo principal,[1] el «protocolo de evaluaciones mediadoras». El método incorpora la mayoría de las estrategias de higiene de las decisiones que hemos introducido en los capítulos anteriores. El protocolo puede aplicarse ampliamente y siempre que la evaluación de un plan o una opción requiera considerar y ponderar múltiples dimensiones. Puede ser utilizado, y adaptado de varias maneras, por organizaciones de todo tipo, incluidas distintas compañías, hospitales, universidades y organismos públicos.

Aquí ilustramos el protocolo con un ejemplo estilizado que es una combinación de varios casos reales: una empresa ficticia que llamaremos Mapco. Seguiremos los pasos que da Mapco al estudiar la oportunidad de realizar una adquisición importante y transformadora, y destacaremos en qué se diferencian de los pasos habituales que da una empresa en una situación semejante. Veremos que las diferencias son significativas (pero sutiles: un observador poco atento podría no advertirlas).

LA PRIMERA REUNIÓN: ACORDAR EL MÉTODO

La idea de adquirir Roadco, un competidor, había estado rondando a Mapco, y había madurado lo suficiente como para que los directivos de la empresa proyectaran una reunión de la junta directiva para discutirla. Joan Morrison, director ejecutivo (CEO) de Mapco, convocó una reunión del comité de estrategia para un debate preliminar sobre la posible adquisición y sobre lo que debería hacerse para mejorar las deliberaciones de la junta a este respecto. Al principio de la reunión, Joan sorprendió al comité con una propuesta: «Me gustaría proponer que probemos un nuevo procedimiento para la reunión de la junta en la que decidiremos sobre la adquisición de

Roadco. El nuevo procedimiento tiene un nombre poco atractivo: “protocolo de evaluaciones mediadoras”, pero la idea es muy sencilla. Se inspira en la similitud entre la evaluación de una opción estratégica y la evaluación de un candidato a un puesto de trabajo.

»Seguro que están familiarizados con las investigaciones que demuestran que las entrevistas estructuradas producen mejores resultados que las no estructuradas, y, más generalmente, con la idea de que estructurar una decisión de contratación la mejora. Saben que nuestro departamento de recursos humanos ha adoptado estos principios para sus decisiones de contratación. Una gran cantidad de investigaciones demuestran que la estructuración de las entrevistas conduce a una precisión mucho mayor (las entrevistas no estructuradas, tal y como las practicábamos, ni siquiera se acercaban).

»Veo una clara similitud entre la evaluación de candidatos y la de opciones en las grandes decisiones: “Las opciones son como los candidatos”. Y esta similitud me lleva a la idea de que debemos adaptar a nuestra tarea, que es la de evaluar opciones estratégicas, el método que funciona para evaluar a los candidatos».

En un principio, los miembros de la comisión se mostraron desconcertados por la comparación. El proceso de contratación, argumentaron, es una máquina bien engrasada que toma numerosas decisiones similares y no está sometida a una gran presión de tiempo. Una decisión estratégica, en cambio, requiere un gran trabajo *ad hoc* y debe tomarse con rapidez. Algunos miembros del comité dejaron claro a Joan que serían hostiles a cualquier propuesta que retrasara la decisión. También les preocupaba aumentar los requerimientos de diligencia debida al personal de investigación de Mapco.

Joan respondió directamente a estas objeciones. Aseguró a sus colegas que el proceso estructurado no retrasaría la decisión. «Se trataría de establecer el orden del día para la reunión de la junta directiva en la que discutiríamos sobre la operación —explicó—. Tendríamos que establecer de antemano una lista de evaluaciones de diferentes aspectos de la operación, del mismo modo que un entrevistador comienza con una descripción del puesto de trabajo que sirve como lista de comprobación de los rasgos o atributos que debe poseer un candidato. Nos aseguraríamos de que la junta discuta estas evaluaciones por separado, una por una, igual que los

entrevistadores en las entrevistas estructuradas evalúan secuencialmente al candidato en aspectos separados. Entonces, y solo entonces, resolveríamos si aceptamos o rechazamos la operación. Este procedimiento será una forma mucho más eficaz de aprovechar la sabiduría colectiva de la junta.

»Si estamos de acuerdo con este método, tendrá, por supuesto, implicaciones sobre cómo debe presentarse la información y sobre cómo ha de trabajar el equipo de negociación para preparar la reunión. Por eso quería conocer su opinión ahora.»

Un miembro del comité, todavía escéptico, preguntó a Joan qué beneficios aportaba la estructura a la calidad de la toma de decisiones en la contratación y por qué creía que esos beneficios se extenderían a una decisión estratégica. Joan le expuso la lógica. El uso del protocolo de evaluaciones mediadoras, explicó, maximiza el valor de la información al mantener las dimensiones de la evaluación independientes entre sí. «Las discusiones que solemos tener en la junta se parecen mucho a las entrevistas no estructuradas —observó—. Tenemos constantemente presente el objetivo final de llegar a una decisión, y procesamos toda la información a la luz de ese objetivo. Empezamos buscando el modo de llegar al acuerdo, y lo alcanzamos tan pronto como podemos. Al igual que un contratador en una entrevista no estructurada, corremos el riesgo de utilizar todo el debate para confirmar nuestras primeras impresiones.

»El uso de un enfoque estructurado nos obligará a posponer el objetivo de tomar una decisión hasta que no hayamos realizado todas las evaluaciones. Asumiríamos las evaluaciones separadas como objetivos intermedios. De este modo, consideraríamos toda la información disponible y nos aseguraríamos de que nuestra conclusión sobre un aspecto de la operación no cambie nuestra opinión sobre otro no relacionado.»

Los miembros del comité acordaron probar el enfoque, pero preguntaron cuáles serían las valoraciones mediadoras. ¿Había una lista de comprobación predefinida que Joan tenía en su mente? «No —respondió ella—. Eso podría ser así si aplicáramos el protocolo a una decisión rutinaria, pero, en este caso, tendríamos que definir nosotros mismos las evaluaciones mediadoras. Esto es de vital importancia: decidir los principales aspectos de la operación que deben evaluarse depende de nosotros.» El comité de estrategia acordó volver a reunirse al día siguiente para hacerlo.

LA SEGUNDA REUNIÓN: DEFINICIÓN DE LAS EVALUACIONES MEDIADORAS

«Lo primero que vamos a hacer —explicó Joan— es confeccionar una lista exhaustiva de evaluaciones independientes sobre la operación. Estas serán evaluadas por el equipo de investigación de Jeff Schneider. Nuestra tarea hoy es elaborar la lista de evaluaciones. Esta debe ser exhaustiva en el sentido de que cualquier hecho relevante que se les ocurra debe tener un sitio e influir en al menos una de las evaluaciones. Y lo que quiero decir con "independiente" es que un hecho relevante debe influir preferentemente solo en una de las evaluaciones para así minimizar la redundancia.»

El grupo se puso a trabajar y elaboró una larga lista de hechos y datos que parecían relevantes. A continuación, los organizó en una lista de evaluaciones. El reto, descubrieron pronto los participantes, era hacer una lista corta, completa y compuesta por evaluaciones que no se solaparan. Sin embargo, la tarea era factible. De hecho, la lista final de siete evaluaciones del grupo era por encima similar a la tabla de contenidos que la junta esperaría de un informe normal que presenta una propuesta de adquisición. Además del modelo financiero previsto, la lista incluía, por ejemplo, una evaluación de la calidad del equipo directivo de la empresa objetivo y una valoración de la probabilidad de que se consiguieran las sinergias previstas.

Algunos miembros del comité de estrategia se mostraron decepcionados por el hecho de que la reunión no produjera nuevas ideas sobre Roadco. Pero, explicó Joan, ese no era el objetivo. El objetivo inmediato era dar instrucciones al equipo encargado de estudiar la adquisición. Cada evaluación, dijo, será objeto de un apartado diferente en el informe del equipo de la operación y se debatirá por separado en la junta.

La misión del equipo de la operación, tal y como la veía Joan, no era decir a la junta lo que pensaba de la operación en su conjunto; al menos, no todavía. Debía proporcionar una evaluación objetiva e independiente sobre cada una de las evaluaciones mediadoras. En última instancia, explicó Joan, cada apartado del informe del equipo de la operación debe finalizar con una evaluación que responda a una simple pregunta: «Dejando a un lado el peso que debamos dar a este tema en la decisión final, ¿con qué fuerza se manifiestan las pruebas de esta evaluación a favor o en contra de la operación?».

El jefe del equipo encargado de evaluar la operación, Jeff Schneider, reunió a su equipo esa tarde para organizar el trabajo. Los cambios con respecto a la forma habitual de trabajar del equipo no fueron muchos, pero hizo hincapié en su importancia.

«En primer lugar —explicó—, los analistas del equipo deben intentar que sus análisis sean lo más objetivos posible. Las evaluaciones deben basarse en hechos —nada nuevo al respecto—, pero también han de tener, siempre que sea posible, una "visión externa".» Como los miembros del equipo no estaban seguros de lo que quería decir con una «visión externa», Jeff les puso dos ejemplos sacados de las evaluaciones mediadoras que Joan había definido. Para evaluar la probabilidad de que la operación recibiese la aprobación de los reguladores, dijo, tendrían que empezar por averiguar la «tasa-base», esto es, el porcentaje de transacciones comparables que se aprueban. Esta tarea requeriría, a su vez, definir una «clase de referencia» relevante, un grupo de operaciones consideradas suficientemente comparables.

A continuación, Jeff explicó cómo evaluar las competencias tecnológicas del departamento de desarrollo de productos de la empresa objetivo, otra evaluación importante que Joan había enumerado. «No basta con describir los logros recientes de la empresa basándose en hechos y llamarlos "buenos" o "excelentes". Lo que espero es algo así como que "este departamento de desarrollo de productos está en el segundo quintil de su grupo de empresas similares, medido por su historial reciente de lanzamientos de productos".» En general, explicó, el objetivo era hacer evaluaciones lo más comparativas posible, porque los juicios relativos son mejores que los absolutos.

Jeff tenía otra petición que hacer. Según las instrucciones de Joan, las evaluaciones deben ser lo más independientes posible entre sí, para reducir de ese modo el riesgo de que una evaluación influya en las demás. En consecuencia, asignó diferentes analistas a las distintas evaluaciones y los instruyó para que trabajaran de forma independiente.

Algunos analistas manifestaron su sorpresa. «¿No es mejor el trabajo en equipo? —le preguntaron— ¿De qué sirve reunir un equipo si no quiere que nos comuniquemos?»

Jeff se dio cuenta de que tenía que explicar la necesidad de la independencia. «Probablemente conozcan el "efecto halo" en la contratación —dijo—. Eso es lo que ocurre cuando la impresión general de un candidato influye en su evaluación de las capacidades del candidato en una faceta específica. Eso es lo que intentamos evitar.» Como algunos de los analistas parecían pensar que ese efecto no era un problema serio, Jeff utilizó otra comparación: «Si tienen cuatro testigos de un delito, ¿dejarían que hablaran entre ellos antes de testificar? Obviamente, no. No querrán que un testigo influya en los demás». Los analistas no encontraron la comparación muy afortunada, pero Jeff pensó que habían captado el mensaje.

En realidad, Jeff no tenía suficientes analistas para lograr el objetivo de unas evaluaciones completamente independientes. Jane, un miembro experimentado del equipo, se encargó de dos evaluaciones. Jeff eligió las dos de modo que fueran lo más distintas posible entre sí, e indicó a Jane que hiciera la primera evaluación y preparara el informe sobre ella antes de pasar a la otra. Otra preocupación era la evaluación de la calidad del equipo; a Jeff le preocupaba que sus analistas tuvieran dificultades para disociar su evaluación de la calidad intrínseca del equipo de los juicios sobre los resultados recientes de la empresa (que el equipo, por supuesto, estudiaría en detalle). Para solucionar este problema, Jeff pidió a un experto externo en recursos humanos que opinara sobre la calidad del equipo directivo. De este modo, pensó, obtendría un informe más independiente.

Jeff tenía otra instrucción que el equipo encontró un tanto inusual. Cada apartado debía centrarse en una evaluación y, como pedía Joan, conducir a una conclusión en forma de calificación. Aunque, añadió Jeff, los analistas tenían que incluir en cada apartado toda la información relevante y basada en los hechos sobre la evaluación. «No ocultéis nada —les indicó—. El tono general del apartado será coherente con la calificación propuesta, por supuesto, pero, si hay información que parece incoherente o incluso contradictoria con la calificación principal, no deben esconder nada bajo la alfombra. Su trabajo no es vender su recomendación. Es representar la verdad. Si es complicado, que lo sea; a menudo lo es.»

En ese mismo espíritu, Jeff animó a los analistas a ser transparentes sobre su nivel de confianza en cada evaluación. «La junta sabe que ustedes no tienen toda la información; la ayudarán si le dicen cuándo realmente tantean

en la oscuridad. Y, si se encuentran con algo que verdaderamente les haga dudar —algo que pudiera romper la operación—, deben, por supuesto, darlo a conocer de inmediato.»

El equipo de la operación procedió como se le indicó. Por fortuna, no encontró ningún escollo importante. Elaboró un informe para Joan y la junta que abarcaba todas las evaluaciones bien identificadas.

LA REUNIÓN DE LA DECISIÓN

Cuando leyó el informe del equipo para preparar la reunión de la decisión, Joan se percató al momento de algo importante: aunque la mayoría de las evaluaciones apoyaban la operación, no pintaban un panorama sencillo, de color de rosa y con todos los sistemas en funcionamiento. Algunas valoraciones eran sólidas; otras, no. Sabía que estas diferencias eran el resultado previsible de mantener las evaluaciones independientes unas de otras. Cuando la coherencia excesiva se mantiene a raya, la realidad no es tan coherente como la mayoría de las presentaciones a la junta hacen parecer. «Bien», pensó Joan. «Estas discrepancias entre las evaluaciones suscitarán preguntas y generarán discusiones. Eso es justo lo que necesitamos para tener un buen debate en la junta. Los distintos resultados no facilitarán la decisión, seguro, pero la harán mejor.»

Joan convocó una reunión de la junta para examinar el informe y tomar una decisión. Explicó el método que siguió el equipo de la operación e invitó a los miembros de la junta a aplicar el mismo principio. «Jeff y su equipo han trabajado mucho para que las evaluaciones sean independientes unas de otras —dijo— y nuestra tarea ahora es examinarlas también de forma independiente. Esto significa que consideraremos cada evaluación por separado, antes de empezar a discutir la decisión final. Vamos a tratar cada evaluación como un punto distinto del orden del día.»

Los miembros de la junta sabían que emplear este método estructurado sería difícil. Joan les pedía que no se formaran una opinión global del acuerdo antes de que se discutieran todas las valoraciones, pero muchos de ellos conocían el sector industrial. «Tenían» una opinión sobre Roadco. No discutirla parecería un poco artificial. Sin embargo, como entendieron lo que Joan intentaba conseguir, aceptaron seguir sus reglas y abstenerse de manera provisional de discutir sus opiniones generales.

Para su sorpresa, los miembros de la junta directiva descubrieron que esta práctica era muy valiosa. Durante la reunión, algunos de ellos incluso cambiaron de opinión sobre la operación (aunque nadie lo sabría nunca, ya que se habían guardado sus opiniones). La forma en que Joan dirigió la reunión desempeñó un papel importante: utilizó el método de «estimar-hablar-estimar», que combina las ventajas de la deliberación con las de promediar opiniones independientes.[2]

Así es como Joan procedió. En cada evaluación, Jeff resumió brevemente en nombre del equipo de la operación los hechos clave (que los miembros de la junta directiva habían leído con detalle de antemano). A continuación, Joan pidió a los miembros de la junta que utilizaran una aplicación de votación en sus teléfonos para dar su propia calificación a la evaluación, ya fuera la misma que la propuesta por el equipo de la operación u otra diferente. La distribución de las calificaciones se proyectó inmediatamente en la pantalla, sin identificar a los calificadores. «Esto no es una votación —explicó Joan—. Solo estamos tomando la temperatura de la sala sobre cada tema.» Al obtener una lectura inmediata de la opinión independiente de cada miembro del consejo antes de iniciar un debate, Joan redujo el peligro de la influencia social y las cascadas de información.

En algunas evaluaciones hubo un consenso inmediato, pero, en otras, el proceso reveló opiniones opuestas. Naturalmente, Joan manejó la discusión para dedicar más tiempo a las últimas. Se aseguró de que los miembros de la junta de cada lado de esa división hablaran animándoles a expresar sus puntos de vista con hechos y argumentos, pero también con matices y humildad. En una ocasión, un entusiasmado miembro de la junta dijo creer con firmeza en la operación, y ella le recordó que «todos somos personas razonables y podemos no estar de acuerdo, así que este debe ser un tema en el que las personas razonables pueden estar en desacuerdo».

Cuando el debate en torno a una evaluación llegaba a su fin, Joan pidió a los miembros de la junta que volvieran a votar con una calificación. La mayoría de las veces hubo más convergencia que en la ronda inicial. La misma secuencia —una primera estimación, una discusión y una segunda estimación— se repitió para cada evaluación.

Finalmente, llegó el momento de llegar a una conclusión sobre la operación. Para facilitar el debate, Jeff mostró la lista de evaluaciones en la pizarra, con la media de las calificaciones que la junta había dado a cada

evaluación. Los miembros de la junta examinaban el perfil de la operación. ¿Qué debían decidir?

Un miembro de la junta hizo una sencilla sugerencia: utilizar una media aritmética de las calificaciones. (Tal vez conocía la superioridad de la suma mecánica sobre el juicio holístico y clínico, como se comenta en el capítulo 9.) Sin embargo, otro miembro objetó de inmediato que, en su opinión, algunas de las evaluaciones deberían tener un peso mucho mayor que otras. Una tercera persona se mostró en desacuerdo y sugirió una jerarquía diferente para las evaluaciones.

Joan interrumpió la discusión. «No se trata de hacer un cálculo, una simple combinación de las calificaciones resultantes de evaluación —dijo —. Hemos retrasado la intuición, pero ahora es el momento de utilizarla. Lo que necesitamos ahora es su juicio.»

Joan no explicó su lógica, pero había aprendido esta lección viendo determinadas prácticas. Sabía que, sobre todo cuando se trata de decisiones importantes, la gente rechaza los planes que les atan las manos y no les dejan usar su juicio. Había visto cómo los responsables de la toma de decisiones juegan con el sistema cuando saben que se utilizará una fórmula. Cambian las calificaciones para llegar a la conclusión deseada, lo que anula el propósito de todo el ejercicio. Además, aunque no fue el caso aquí, se mantuvo alerta ante la posibilidad de que surgieran consideraciones decisivas que no se habían previsto en la definición de las evaluaciones (los factores «de pierna rota» de los que hablamos en el capítulo 10). Si aparecieran estos factores imprevistos que rompen la operación (o, a la inversa, la aprueban), un proceso de decisión puramente mecánico basado en la media de las evaluaciones podría conducir a un grave error.

Joan también sabía que dejar que los miembros de la junta utilizaran su intuición en esta fase era muy diferente a que lo hicieran en otra anterior del proceso. Ahora que las evaluaciones estaban disponibles y eran conocidas por todos, la decisión final estaba anclada con seguridad en esas calificaciones basadas en hechos y discutidas a fondo. Un miembro de la junta tendría que dar razones de peso para estar en contra de la operación mientras mirase una lista de evaluaciones mediadoras que la apoyaran en su mayoría. Siguiendo esta lógica, la junta debatió sobre la operación y la votó tal como se hace en todas las juntas directivas.

EL PROTOCOLO DE EVALUACIONES MEDIADORAS EN DECISIONES RECURRENTES

Hemos descrito el protocolo de evaluaciones mediadoras en el contexto de una decisión única y singular. Sin embargo, el procedimiento es también aplicable a las decisiones recurrentes. Imaginemos que Mapco no realiza una única adquisición, sino que es un fondo de capital de riesgo que realiza repetidas inversiones en empresas de nueva creación. El protocolo sería igualmente aplicable, y la historia sería casi la misma, con solo dos giros que, si acaso, la hacen más sencilla.

En primer lugar, el paso inicial —definir la lista de evaluaciones mediadoras— solo debe hacerse una vez. El fondo tiene criterios de inversión que aplica a todas sus posibles inversiones: son las evaluaciones. No es necesario reinventarlos cada vez.

En segundo lugar, si el fondo toma muchas decisiones del mismo tipo, puede utilizar su experiencia para calibrar sus juicios. Consideremos, por ejemplo, una evaluación que todo fondo querrá hacer: evaluar la calidad del equipo de gestión. Ya sugerimos que estas evaluaciones debían realizarse en relación con una clase de referencia. Tal vez el lector se compadeciera de los analistas de Mapco: reunir datos sobre empresas comparables, además de evaluar un objetivo específico, es todo un reto.

Los juicios comparativos resultan mucho más fáciles en el contexto de una decisión recurrente. Si se han evaluado los equipos gestores de decenas, incluso cientos de empresas, se puede utilizar esa experiencia compartida como una clase de referencia. Una forma práctica de hacerlo es crear una escala de casos definida por casos de anclaje. Se puede decir, por ejemplo, que el equipo gestor de la empresa objetivo es «tan bueno como el equipo gestor de la empresa ABC cuando la adquirimos», pero no tanto que es «tan bueno como el equipo gestor de la empresa DEF». Por supuesto, los casos de anclaje deben ser conocidos por todos los participantes (y actualizados de forma periódica). Definirlos requiere una inversión inicial de tiempo. Pero el valor de este enfoque es que los juicios relativos (comparar ese equipo con los de ABC y DEF) son mucho más fiables que las calificaciones absolutas en una escala definida por números o adjetivos.

QUÉ COSAS CAMBIA EL PROTOCOLO

Para facilitar la comprensión, resumimos en la tabla 4 los principales cambios que implica el protocolo de evaluaciones mediadoras:

1. Al comienzo del proceso, habrá que estructurar la decisión en valoraciones mediadoras. (*En el caso de los juicios recurrentes, esto se hará una sola vez*.)
2. Habrá que asegurarse de que, siempre que sea posible, las evaluaciones mediadoras tengan una visión externa. (*En el caso de los juicios recurrentes, se utilizarán juicios relativos, con una escala de casos si es posible*.)
3. En la fase analítica, habrá que mantener las evaluaciones tan independientes entre sí como sea posible.
4. En la reunión de la decisión, habrá que examinar cada evaluación por separado.
5. En cada evaluación, habrá que asegurarse de que los participantes emitan sus juicios de forma individual; a continuación, se ha de utilizar el método de estimar-hablar-estimar.
6. Para tomar la decisión final, habrá que retrasar la intuición, pero no excluirla.

TABLA 4. Principales pasos del protocolo de evaluaciones mediadoras

Es posible que el lector haya reconocido aquí una aplicación de varias de las técnicas de higiene de las decisiones que hemos presentado en los capítulos anteriores: secuenciar la información, estructurar la decisión en evaluaciones independientes, utilizar un marco de referencia común basado en la visión externa y sumar los juicios independientes de numerosos individuos. Con la aplicación de estas técnicas, el protocolo de evaluaciones mediadoras aspira a cambiar el «proceso» de decisión para introducir la mayor higiene posible en las decisiones.

Este énfasis en el proceso, en contraposición con el contenido de las decisiones, sin duda descolocará a muchos. Las reacciones de los miembros del equipo de investigación y de los miembros de la junta directiva, tal y como las hemos descrito, no son excepcionales. El contenido es específico;

el proceso es genérico. Utilizar la intuición y el juicio es divertido; seguir el proceso no lo es. La sabiduría convencional sostiene que las buenas decisiones —sobre todo las mejores— nacen de la perspicacia y la creatividad de los grandes líderes. (Nos gusta creerlo especialmente cuando somos el líder en cuestión.) Y, para muchos, la palabra «proceso» evoca la burocracia, el papeleo y los retrasos.

Nuestra experiencia con empresas y organismos públicos que han aplicado todos o algunos de los componentes del protocolo indica que estas preocupaciones son desacertadas. Sin duda, añadir complejidad a los procesos de toma de decisiones de una organización que ya es burocrática no mejorará las cosas. Pero la higiene de las decisiones no tiene por qué ser lenta y, desde luego, no tiene por qué ser burocrática. Por el contrario, promueve el reto y el debate, no el consenso asfixiante que caracteriza a las burocracias.

Los argumentos en favor de la higiene de las decisiones son claros. Los dirigentes de las empresas y del sector público suelen ignorar por completo el ruido en sus decisiones más importantes. Por ello, no toman ninguna medida específica para reducirlo. En ese sentido, son como los contratadores que siguen confiando en las entrevistas no estructuradas como única herramienta de selección de personal: ajenos al ruido en su propio juicio, más confiados en su validez de lo que deberían y sin conocer los procedimientos que podrían mejorarlo.

El lavado de manos no previene contra todas las enfermedades. Y la higiene de las decisiones no evitará todos los errores. No hará que todas las decisiones sean brillantes. Pero, como el lavado de manos, aborda un problema invisible, aunque generalizado y perjudicial. Dondequiera que haya juicio, hay ruido, y proponemos la higiene de las decisiones como herramienta para reducirlo.

SOBRE EL PROTOCOLO DE EVALUACIONES MEDIADORAS

- «Tenemos un proceso estructurado para tomar decisiones de contratación. ¿Por qué no tenemos uno para las decisiones estratégicas? Después de todo, las opciones son como los candidatos.»

- «Es una decisión difícil. ¿Cuáles son las evaluaciones mediadoras en las que debe basarse?»
- «Nuestro juicio intuitivo y holístico sobre este plan es muy importante, pero no lo discutamos todavía. Nuestra intuición nos servirá de mucho más cuando esté informada por las evaluaciones separadas que hemos pedido.»

SEXTA PARTE

Ruido óptimo

En 1973, el juez Marvin Frankel tenía razón al pedir un esfuerzo continuado para reducir el ruido en las sentencias penales. Su auditoría informal e intuitiva del ruido, seguida de intentos más formales y sistemáticos, descubrió desigualdades injustificadas en la forma de tratar a personas iguales. Aquellas desigualdades eran intolerables, e incluso alarmantes.

Gran parte de este libro puede entenderse como un intento de generalizar los argumentos de Frankel y ofrecer una explicación de sus fundamentos psicológicos. Para algunas personas, el ruido en el sistema de justicia penal es de todo punto intolerable, incluso escandaloso. Sin embargo, en otros innumerables contextos tampoco es tolerable, ya que personas supuestamente intercambiables en los sectores privado y público emiten juicios diferentes en el ejercicio de su profesión. En las compañías de seguros, en la contratación y evaluación de empleados, en la medicina, en la ciencia forense, en la educación, en las empresas y en los gobiernos, el ruido interpersonal es una fuente de abundantes errores. Asimismo hemos visto que cada uno de nosotros está sujeto al «ruido de ocasión», en el sentido de que factores en apariencia irrelevantes pueden llevarnos a emitir juicios distintos por la mañana y por la tarde, o los lunes y los jueves.

Pero, como revela la reacción judicial francamente negativa a las directrices para la imposición de penas, los intentos de reducción del ruido suelen tropezar con objeciones serias e incluso apasionadas. Muchas personas han argumentado que las directrices son rígidas, deshumanizadoras y, a su manera, injustas. Casi todo el mundo ha tenido la experiencia de hacer una solicitud razonable a una empresa, un empleador o al Gobierno, solo para encontrarse con esta respuesta: «Nos gustaría ayudarle, pero tenemos las manos atadas. Seguimos normas estrictas». Las normas en cuestión pueden parecer estúpidas e incluso crueles, pero es posible que se hayan adoptado por una buena razón: reducir el ruido (y quizá también el sesgo).

Aun así, algunos intentos de reducir el ruido provocan serias preocupaciones, sobre todo si dificultan o impiden que a los ciudadanos se les escuche como merecen. El uso de algoritmos y del aprendizaje

automático ha puesto esta objeción bajo una nueva luz. Nadie camina con una pancarta que diga: «¡Algoritmos ya!».

Una crítica muy influyente fue la que hicieron Kate Stith, de la facultad de Derecho de Yale, y José Cabranes, juez federal. Lanzaron un enérgico ataque a las directrices para las sentencias y, en cierto modo, a uno de nuestros principales argumentos. Su argumentación se limitaba al ámbito de las sentencias penales, pero puede usarse como objeción a muchas estrategias de reducción del ruido en la educación, las empresas, los deportes y cualquier otro terreno. Stith y Cabranes sostenían que las directrices para la imposición de penas estaban motivadas «por el miedo al ejercicio de la discrecionalidad —por el miedo a juzgar— y por una fe tecnocrática en los expertos y en la planificación central». Argumentaban que el «miedo a juzgar» coopera para prohibir la consideración de «las particularidades de cada caso juzgado». En su opinión, «ninguna solución mecánica puede satisfacer las demandas de justicia».[1]

Merece la pena examinar estas objeciones. En entornos que implican juicios de todo tipo, la gente suele considerar que las «demandas de la justicia» prohíben cualquier tipo de solución mecánica; y, por tanto, permiten, o incluso hacen obligatorio, el uso de procedimientos que son una garantía de ruido. Muchos piden que se preste atención a «las particularidades de cada caso». En hospitales, escuelas y grandes y pequeñas empresas, este llamamiento tiene un profundo atractivo intuitivo. Hemos visto que la higiene de las decisiones incluye distintas estrategias para reducir el ruido, y la mayoría de ellas no implican soluciones mecánicas; cuando las personas descomponen un problema en las partes que lo integran, sus juicios no tienen por qué ser mecánicos. Aun así, muchas personas no verían con buenos ojos el uso de estrategias de higiene en la toma de decisiones.

Hemos definido el ruido como «variabilidad no deseada», y, si no se desea algo, se supone que debe ser eliminado. Sin embargo, el análisis es más complicado y más interesante que eso. El ruido puede ser indeseado en igualdad de condiciones. Pero otras cosas podrían no ser iguales, y los costes de eliminar el ruido podrían superar los beneficios. Hasta cuando un análisis de costes y beneficios sugiere que el ruido es costoso, su eliminación podría producir una variedad de consecuencias negativas o incluso inaceptables para las instituciones públicas y privadas.

Hay siete objeciones principales a los intentos de reducir o eliminar el ruido.

La primera señala que reducir el ruido puede ser caro; podría no merecer la pena. Las medidas necesarias para reducir el ruido podrían resultar muy gravosas. En algunos casos, ni siquiera serían factibles.

La segunda sostiene que algunas estrategias introducidas para reducir el ruido podrían introducir errores propios. En ocasiones, podrían producir un sesgo sistemático. Si todos los analistas de una oficina pública adoptaran los mismos supuestos demasiado optimistas, sus predicciones no serían ruidosas, pero sí estarían equivocadas. Si todos los médicos de un hospital recetaran aspirinas para todas las enfermedades, no habría ruido, pero sí innumerables errores.

Analizaremos estas objeciones en el capítulo 26. Y, en el capítulo 27, nos ocuparemos de cinco objeciones más, que también son comunes y probablemente se oigan en muchos ámbitos en los próximos años, sobre todo con la creciente confianza en las reglas, los algoritmos y el aprendizaje automático.

La tercera objeción sostiene que, si queremos que la gente se sienta tratada con respeto y dignidad, quizá debamos tolerar algo de ruido. Este puede ser un subproducto de un proceso limitado que la gente acaba aceptando porque el proceso ofrece a todos (empleados, clientes, solicitantes, estudiantes, acusados de delitos) una audiencia individualizada, una oportunidad de influir en el ejercicio de la discrecionalidad y la sensación de que han tenido la oportunidad de ser vistos y escuchados.

La cuarta objeción es que el ruido podría ser esencial para dar cabida a nuevos valores y, por tanto, permitir la evolución moral y política. Si eliminamos el ruido, podríamos reducir nuestra capacidad de respuesta cuando los compromisos morales y políticos se mueven en direcciones nuevas e inesperadas. Un sistema sin ruido podría congelar los valores existentes.

La quinta objeción sostiene que algunas estrategias diseñadas para reducir el ruido podrían impulsar el comportamiento oportunista al permitir a la gente engañar al sistema o eludir las prohibiciones. Un poco de ruido, o tal vez mucho, puede ser necesario para evitar ciertos males.

La sexta objeción asegura que un proceso ruidoso puede ser un buen elemento disuasorio. Si la gente sabe que puede ser objeto de una pequeña o

una gran sanción, evitará cometer infracciones, al menos si tiene aversión al riesgo. Un sistema podría tolerar el ruido como forma de producir una disuasión añadida.

Y la séptima objeción remarca que las personas no quieren ser tratadas como si fueran cosas o engranajes de algún tipo de máquina. Ciertas estrategias de reducción del ruido pueden aplastar la creatividad y resultar desmoralizantes.

Aunque abordaremos estas objeciones con la mayor comprensión posible, no las respaldaremos en absoluto, al menos si se toman como razones para rechazar el objetivo general de reducir el ruido. Presagiando la recurrencia de una cuestión a lo largo de nuestro análisis, diremos que el que una objeción sea convincente depende de la estrategia concreta de reducción del ruido que se pretende aplicar. Por ejemplo, alguien podría oponerse a las directrices rígidas y, al mismo tiempo, estar de acuerdo en que la suma de juicios independientes es una buena idea. Podría oponerse al uso del protocolo de evaluaciones mediadoras y estar muy a favor del uso de una escala compartida basada en una visión externa. Teniendo en cuenta estos puntos, nuestra conclusión general es que, aunque las objeciones sean respetables, la reducción del ruido sigue siendo un objetivo digno y hasta urgente. En el capítulo 28 defenderemos esta conclusión analizando un dilema al que las personas se enfrentan cada día, aunque no siempre tengan conciencia de él.

26
Los costes de reducir el ruido

Cuando a una organización se le pide que elimine el ruido, puede objetar que las medidas necesarias son demasiado caras. En circunstancias extremas, la reducción del ruido simplemente no es posible. Hemos escuchado esta objeción en las empresas, en la educación, en el Gobierno y en otros ámbitos. La preocupación es legítima, pero se exagera con facilidad y, a menudo, es solo una excusa.

Para poner la objeción bajo su luz más atractiva, consideremos el caso de un profesor que califica veinticinco redacciones de alumnos de instituto cada semana del curso escolar. Si el profesor no dedica más de quince minutos a cada redacción, la calificación puede ser ruidosa y, por tanto, inexacta e injusta. El profesor podría considerar un poco de higiene en la toma de sus decisiones, tal vez pidiendo a un colega que califique también las redacciones, de modo que, al haber dos personas que lean cada trabajo, el ruido se reduzca. Tal vez el profesor podría lograr el mismo objetivo dedicando más tiempo a la lectura de cada redacción, estructurando el proceso relativamente complejo de la evaluación o leyendo las redacciones más de una vez y en diferentes órdenes. Una detallada directriz de calificación utilizada como lista de control podría ayudar. El profesor también podría asegurarse de leer cada redacción a la misma hora del día para reducir el ruido de ocasión.

Sin embargo, si los juicios del profesor son bastante certeros y no demasiado ruidosos, sería sensato no hacer ninguna de estas cosas. Quizá no mereciera la pena. El profesor podría pensar que utilizar una lista de control o pedir a un colega que lea los mismos trabajos sería una exageración. Para saber si lo es, se necesitaría un análisis sistemático: ¿cuánta más precisión obtendría el profesor? ¿Qué importancia tiene un mayor acierto y cuánto tiempo y dinero requeriría el esfuerzo por reducir el ruido? Podríamos imaginar fácilmente un límite para lo que es necesario invertir en la reducción del ruido. Podríamos ver fácilmente que este límite debería ser distinto cuando los trabajos los escriben estudiantes de un curso

inferior o cuando se trata de trabajos de fin de curso, donde la admisión en la universidad puede estar en juego y la responsabilidad es mayor.

El análisis básico podría ampliarse a situaciones más complejas relacionadas con organizaciones privadas y públicas de todo tipo, que rechazarían algunas estrategias de reducción del ruido. En el caso de algunas enfermedades, los hospitales y los médicos pueden esforzarse por definir pautas sencillas para eliminar la variabilidad. En el caso de los diagnósticos médicos diferentes, los intentos de reducir el ruido tienen un particular interés: podrían salvar vidas. Sin embargo, hay que tener en cuenta la viabilidad y los costes de esos intentos. Una prueba podría eliminar el ruido en los diagnósticos, pero, si la prueba es invasiva, peligrosa y costosa, y si la variabilidad en los diagnósticos es limitada y solo tiene consecuencias leves, entonces puede que no merezca la pena que todos los médicos pidan a todos los pacientes que se sometan a dicha prueba.

Raro es que la evaluación de empleados sea una cuestión de vida o muerte, pero el ruido puede resultar injusto para los empleados y suponer un alto coste para la empresa. Hemos visto que los intentos de reducir el ruido deben ser viables. ¿Merecen la pena? Los casos que implican evaluaciones claramente erróneas pueden llamar la atención y parecer embarazosos, vergonzosos o algo peor. Sin embargo, una institución puede pensar que no vale la pena tomar medidas correctivas detalladas. A veces esa conclusión es miope, interesada y errónea, incluso catastrófica. Quizá mereciera la pena algún tipo de higiene de las decisiones. Sin embargo, la creencia de que reducir el ruido sale demasiado caro no siempre está equivocada.

En suma, hay que comparar los beneficios de la reducción del ruido con los costes. Esto es justo, y es una de las razones por las que las auditorías del ruido son tan importantes. En muchas situaciones, las auditorías revelan que el ruido está produciendo niveles escandalosos de injusticia, costes muy elevados o ambas cosas. Si es así, el coste de la reducción del ruido no es una buena razón para no intentarlo.

¿MENOS RUIDO, MÁS ERRORES?

Otra objeción es que algunos intentos de reducir el ruido podrían producir de por sí niveles de error intolerables. La objeción puede ser convincente si los instrumentos utilizados para reducir el ruido son demasiado contundentes. De hecho, algunos intentos de reducir el ruido podrían incluso aumentar el sesgo. Si una plataforma de medios sociales como Facebook o Twitter introdujera directrices firmes que exigieran la eliminación de todas las publicaciones que contengan determinadas palabras vulgares, sus decisiones serían menos ruidosas, pero estarían excluyendo numerosos *posts* que tendrían derecho a ser difundidos. Estos falsos positivos son un error direccional (un sesgo).

La vida está trufada de reformas institucionales destinadas a reducir la discrecionalidad de las personas y las prácticas que generan ruido. Muchas de estas reformas tienen sus razones, pero algunos remedios son peores que la enfermedad. En *The Rhetoric of Reaction*, el economista Albert Hirschman señala tres objeciones comunes a los propósitos de reforma. En primer lugar, esos propósitos podrían ser contumaces, en el sentido de que agravarían el propio problema que pretenden resolver.[1] En segundo lugar, podrían ser inútiles; podrían no cambiar las cosas en absoluto. Y, en tercer lugar, ponen en peligro otros valores importantes (como cuando se dice que un intento de proteger a los sindicatos y el derecho a sindicarse perjudica el crecimiento económico). La contumacia, la inutilidad y el peligro pueden ofrecerse como objeciones a la reducción del ruido, y de las tres, las de la contumacia y el peligro tienden a ser las más poderosas. A veces, estas objeciones no son más que retórica, un esfuerzo por hacer descarrilar una reforma que en realidad hará mucho bien. Sin embargo, algunas estrategias de reducción del ruido podrían poner en peligro valores importantes, y, en otras, el riesgo de contumacia podría no ser fácilmente descartado.

Los jueces que se opusieron a las directrices de las sentencias señalaban ese riesgo. Conocían bien el trabajo del juez Frankel, y no negaron que la discrecionalidad genera ruido. Sin embargo, pensaron que reducir la discrecionalidad produciría más errores, no menos. Citando a Václav Havel, insistieron: «Tenemos que abandonar la arrogante creencia de que el mundo no es más que un rompecabezas que hay que resolver, una máquina con instrucciones de uso que esperan ser descubiertas, un cuerpo de información que hay que introducir en un ordenador con la esperanza de que, tarde o temprano, escupa una solución universal».[2] Una de las razones para

rechazar la idea de las soluciones universales es la pertinaz creencia de que las situaciones humanas son muy variadas y de que los buenos jueces tienen presentes dichas variaciones (lo que significaría tolerar el ruido o, al menos, rechazar algunas estrategias de reducción de ruido).

En los primeros tiempos del ajedrez por ordenador, una gran compañía aérea ofrecía un programa de ajedrez para los pasajeros internacionales, a los que se invitaba a jugar contra un ordenador. El programa tenía varios niveles. En el más bajo, el programa utilizaba una regla sencilla: poner en jaque al rey del oponente siempre que se pueda. El programa no era ruidoso. Jugaba siempre de la misma manera; siempre seguía su sencilla regla. Sin embargo, la regla arrojaba una gran cantidad de errores. El programa era espantoso en el juego. Incluso jugadores inexpertos podían ganar (lo que sin duda era el objetivo; los pasajeros ganadores son pasajeros contentos).

Otro ejemplo es la política de sentencias penales llamada «three strikes and you’re out», adoptada en algunos estados de Estados Unidos.[3] La idea es que, quien cometa tres delitos graves, tendrá cadena perpetua (y punto). Esta política reduce la variabilidad resultante de la asignación aleatoria del juez que emita la sentencia. Algunos de sus defensores estaban especialmente preocupados por el ruido de nivel y la posibilidad de que algunos jueces fueran demasiado indulgentes con los delincuentes reincidentes. La eliminación del ruido es el punto central de la legislación de los «three strikes».

Pero, aunque la política de los «three strikes» consiga su objetivo de reducir el ruido, podemos objetar razonablemente que el precio de ese logro es demasiado alto. Algunas personas que han cometido tres delitos no deberían ser encerradas de por vida. Quizá sus delitos no fueran violentos. O las terribles circunstancias de su vida contribuyeran a llevarlos a la delincuencia. Tal vez muestren alguna aptitud para la rehabilitación. Mucha gente piensa que la cadena perpetua, sin tener en cuenta las circunstancias particulares, no solo es demasiado dura, sino también intolerablemente rígida. Por eso, el precio de esta estrategia de reducción del ruido es altísimo.

Consideremos el caso de «Woodson contra Carolina del Norte», en el que el Tribunal Supremo de Estados Unidos sostuvo que una pena capital obligatoria no era inconstitucional porque fuera demasiado brutal, sino

«porque era una norma».[4] El objetivo de la pena capital obligatoria era garantizar que no hubiera ruido (hay que aclarar que, en circunstancias determinadas, los asesinos debían ser condenados a la pena capital). Invocando la necesidad de un tratamiento individualizado, el tribunal dijo que «ya no prevalece la creencia de que para cada delito que entre en una categoría legal similar debe haber un castigo idéntico, sin tener en cuenta la vida pasada y los hábitos del delincuente». Según el Tribunal Supremo, un grave defecto constitucional de la pena capital obligatoria es que «no trata a todas las personas condenadas por un delito definido como seres humanos individuales y singulares, sino como miembros de una masa indiferenciada y sin rostro que debe ser sometida a la imposición ciega de la pena capital».

La pena capital entraña un riesgo muy elevado, por supuesto, pero el análisis del tribunal puede aplicarse a muchas otras situaciones, la mayoría de ellas no relacionadas con el derecho. Los profesores que evalúan a sus alumnos, los médicos que evalúan a los pacientes, los contratadores que evalúan a los empleados, los aseguradores que fijan las primas, los entrenadores que evalúan a los deportistas... Todas estas personas podrían cometer errores si aplicaran normas demasiado rígidas que redujeran el ruido. Si los contratadores utilizan reglas sencillas para evaluar, ascender o suspender a los empleados, esas reglas podrían eliminar el ruido y dejar de lado aspectos importantes del rendimiento de los empleados. Un sistema de puntuación sin ruido que no tenga en cuenta variables significativas podría ser peor que confiar en (ruidosos) juicios individuales.

En el capítulo 27 se tiene en cuenta la idea general de tratar a las personas como «individuos únicos», en lugar de como «miembros de una masa indiferenciada y sin rostro». Por ahora, nos centraremos en un punto más prosaico. Algunas estrategias de reducción del ruido producen demasiados errores; pueden ser muy parecidas a aquel ingenuo programa de ajedrez por ordenador.

Aun así, la objeción parece mucho más convincente de lo que realmente es. Si una estrategia de reducción del ruido es propensa a los errores, no debemos contentarnos con altos niveles de ruido. Más bien deberíamos tratar de idear una mejor estrategia de reducción del ruido; por ejemplo, sumar juicios en lugar de adoptar reglas ingenuas o establecer directrices acertadas o normas en lugar de aquellas reglas imprudentes. En aras de la reducción del ruido, una universidad podría decir, por ejemplo, que admitirá

a las personas con las notas más altas y nada más. Si esa regla parece demasiado tosca, la universidad podría crear una fórmula que tenga en cuenta las puntuaciones obtenidas en determinados test, las calificaciones medias, la edad, los logros deportivos, el entorno familiar, etcétera. Las reglas complejas podrían ser más precisas, más ajustadas a toda la gama de factores relevantes. Del mismo modo, los médicos tienen normas complejas para diagnosticar algunas enfermedades. Las directrices y normas que utilizan los profesionales no siempre son sencillas o rigurosas, y muchas de ellas ayudan a reducir el ruido sin inducir costes (o sesgos) insufriblemente altos. Y, si las directrices o las normas no funcionan, tal vez podrían introducirse otras formas de higiene de las decisiones adaptadas a situaciones concretas que sí funcionen; recordemos la suma de juicios o el uso de un proceso estructurado como el protocolo de evaluaciones mediadoras.

ALGORITMOS SIN RUIDO Y SESGADOS

Los potencialmente elevados costes de la reducción del ruido aparecen a menudo en el contexto de los algoritmos, donde hay crecientes objeciones al «sesgo algorítmico». Como hemos visto, los algoritmos eliminan el ruido y, a menudo, parecen atractivos por esa razón. De hecho, gran parte de este libro podría tomarse como un argumento a favor de una mayor confianza en los algoritmos simplemente porque no son ruidosos. Sin embargo, como también hemos visto, la reducción del ruido puede tener un coste intolerable si una mayor dependencia de los algoritmos aumenta la discriminación por motivos de raza o sexo, o si afecta a los miembros de grupos desfavorecidos.

Existe el temor generalizado de que los algoritmos tengan de hecho esa consecuencia discriminatoria, lo que sin duda constituye un serio riesgo. En *Weapons of Math Destruction*, la matemática Cathy O'Neil alerta de que la confianza en los *big data* y la decisión por algoritmos puede incrustar en la mente prejuicios, aumentar la desigualdad y amenazar la propia democracia.[5] Otro argumento escéptico es el de que «los modelos matemáticos potencialmente sesgados están alterando nuestra vida, y ni las empresas responsables de su desarrollo ni el Gobierno están interesados en abordar el problema».[6] Según ProPublica, una organización

independiente de periodismo de investigación, COMPAS, un algoritmo ampliamente utilizado en las evaluaciones de riesgo de reincidencia, tiene un fuerte sesgo contra los miembros de las minorías raciales.[7]

Nadie debe dudar de que es posible —incluso fácil— crear un algoritmo que esté libre de ruido pero que sea racista, sexista o con otro tipo de sesgo. Un algoritmo que utilice explícitamente el color de piel de un acusado para determinar si puede recibir la libertad bajo fianza sería discriminatorio (y su uso sería ilegal en muchas naciones). Un algoritmo que tenga en cuenta si las solicitantes de empleo podrían quedarse embarazadas, discriminaría a las mujeres. En estos y otros casos, los algoritmos pueden eliminar la variabilidad no deseada en el juicio, pero también encierran un sesgo intolerable.

En principio, deberíamos ser capaces de diseñar un algoritmo que «no» tenga en cuenta la raza o el sexo. De hecho, un algoritmo podría estar diseñado sin tener en cuenta estas cuestiones. El mayor problema, que ahora recibe mucha atención, es que un algoritmo podría discriminar, y estar sesgado en este sentido, incluso cuando no utiliza abiertamente la raza y el género como predictores.

Como hemos señalado, un algoritmo podría estar sesgado por dos razones. La primera es que podría estar o no estar diseñado para utilizar predictores que estén altamente correlacionados con la raza o el sexo. Por ejemplo, la estatura y el peso están correlacionados con el sexo, y el lugar donde alguien se crio o donde actualmente vive está correlacionado con la raza. La segunda razón es que la discriminación también puede provenir de la fuente de los datos. Si un algoritmo es entrenado en un conjunto de datos que esté sesgado, también estará sesgado. Considérense los algoritmos de «policía predictiva», que intentan predecir la delincuencia, a menudo con el fin de mejorar la asignación de los recursos policiales.[8] Si los datos existentes sobre la delincuencia reflejan el exceso de vigilancia de ciertos barrios o el exceso de denuncias de ciertos tipos de delitos, los algoritmos resultantes perpetuarán o exacerbarán la discriminación. Siempre que haya un sesgo en los datos de entrenamiento, es muy posible que se diseñe, de forma deliberada o no, un algoritmo que codifique la discriminación. De esto se deduce que, aunque un algoritmo no considere expresamente la raza o el género, podría estar tan sesgado como los seres humanos. En este sentido, los algoritmos podrían ser peores: como eliminan el ruido, podrían

estar más «fidedignamente» sesgados que las mentes de los juzgadores humanos.[9]

Para muchas personas, una consideración práctica clave es si un algoritmo tiene un efecto dispar en grupos identificables. La forma exacta de probar el efecto dispar y decidir qué constituye discriminación, sesgo o equidad en un algoritmo es un asunto sorprendentemente complejo que queda fuera del alcance de este libro.[10]

Sin embargo, el hecho de que se pueda plantear esta cuestión es una clara ventaja de los algoritmos sobre los juicios humanos. Para empezar, recomendamos una cuidadosa evaluación de los algoritmos para garantizar que no consideren *inputs* inadmisibles y comprobar si discriminan de forma objetiva. Mucho más difícil es someter a los seres humanos, cuyos juicios son a menudo opacos, al mismo tipo de escrutinio; las personas a veces discriminan inconscientemente y de maneras que los observadores externos, incluido el sistema jurídico, no pueden advertir fácilmente. Así que, en cierto modo, un algoritmo puede ser más transparente que los seres humanos.

Sin duda, tenemos que llamar la atención sobre los costes de los algoritmos sin ruido pero tendenciosos, al igual que debemos considerar los costes de las reglas sin ruido pero sesgadas. La cuestión clave es si podemos diseñar algoritmos que lo hagan mejor que los juzgadores humanos del mundo real con una combinación de criterios de primordial importancia: exactitud con reducción del ruido, no discriminación y equidad. Muchas pruebas indican que los algoritmos pueden superar a los seres humanos en cualquier combinación de criterios que seleccionemos. (Nótese que decimos «pueden», no «podrán».) Por ejemplo, y como se describe en el capítulo 10, un algoritmo puede ser más exacto que los juzgadores humanos en lo que respecta a las decisiones sobre fianzas y producir menos discriminación racial que los seres humanos. Del mismo modo, un algoritmo de selección de currículos puede elegir un conjunto de talentos mejor «y más variado» que los examinadores humanos de currículos.

Estos ejemplos, y muchos otros, llevan a una conclusión irrefutable: aunque es poco probable que un algoritmo de predicción sea perfecto en un mundo incierto, puede ser mucho menos imperfecto que el ruidoso y, a menudo, sesgado juicio humano. Esta superioridad se da tanto en términos

de validez (los buenos algoritmos casi siempre predicen mejor) como de discriminación (los buenos algoritmos pueden estar menos sesgados que los juzgadores humanos). Si los algoritmos cometen menos errores que los expertos humanos y, sin embargo, tenemos una preferencia intuitiva por las personas, nuestras preferencias intuitivas deberían ser examinadas cuidadosamente.

Nuestras conclusiones generales son sencillas y se extienden mucho más allá de la cuestión de los algoritmos. Es cierto que las estrategias de reducción del ruido pueden ser costosas. Sin embargo, la mayoría de las veces sus costes son una mera excusa, no una razón suficiente, para tolerar la injusticia y los costes del ruido. Por supuesto, los intentos de reducir el ruido pueden ocasionar errores propios, quizá en forma de sesgo. En este caso tenemos un serio problema, pero la solución no es abandonar los intentos de reducir el ruido, sino aumentarlos.

SOBRE LOS COSTES DE LA REDUCCIÓN DEL RUIDO

- «Si intentáramos eliminar el ruido en la educación, tendríamos que gastar mucho dinero. Los profesores son ruidosos cuando califican a los alumnos. No podemos tener cinco profesores calificando el mismo trabajo.»
- «Si, en lugar de confiar en el juicio humano, una red social decide que nadie puede usar ciertas palabras, sea cual fuere el contexto, eliminará el ruido, pero también producirá muchos errores. El remedio podría ser peor que la enfermedad.»
- «Es cierto que hay reglas y algoritmos que tienen sesgos. Sin embargo, las personas también tienen sesgos. Lo que debemos preguntarnos es si podemos diseñar algoritmos que estén libres de ruido y menos sesgados.»
- «Puede ser costoso eliminar el ruido, pero el coste suele merecer la pena. El ruido puede ser tremendamente injusto. Y, si un intento de reducir el ruido es demasiado burdo —si acabamos estableciendo directrices o reglas que son intolerablemente rígidas o que inadvertidamente producen un sesgo—, no debemos rendirnos. Tenemos que volver a intentarlo.»

27
Dignidad

Supongamos que a una persona le han denegado una hipoteca, pero no porque alguien haya estudiado su situación, sino porque un banco tiene la norma inflexible de que las personas con su calificación crediticia simplemente no pueden obtener una hipoteca. Supongamos también que otra persona tiene unas calificaciones inmejorables y que un entrevistador de una empresa quedó muy impresionado con ella, pero su solicitud de empleo es rechazada porque fue condenada por un delito relacionado con las drogas hace quince años y la empresa tiene prohibido contratar a alguien que haya sido condenado por un delito. Imaginemos que a alguien se le acusa de un delito y se le deniega la libertad bajo fianza, no tras una audiencia individualizada ante un ser humano real, sino porque un algoritmo ha decidido que las personas con sus características tienen un riesgo de fuga que supera el límite establecido para la libertad bajo fianza.

En estos casos, muchas personas se opondrían. Quieren ser tratadas como individuos. Quieren que un ser humano real conozca sus circunstancias particulares. Pueden o no ser conscientes de que el tratamiento individualizado producirá ruido. Sin embargo, si ese es el precio de dicho tratamiento, insisten en que es un precio que merece la pena pagar. Pueden quejarse cuando no son tratadas, en palabras del Tribunal Supremo, «como seres humanos individuales y singulares, sino como miembros de una masa indiferenciada y sin rostro que debe ser sometida a la imposición ciega» de una pena (véase el capítulo 26).

Muchos son los que insisten en la necesidad de una audiencia individualizada, libre de lo que consideran la tiranía de las normas, para que las personas tengan la sensación de que se las trata como individuos y, por tanto, con cierto respeto. La idea del debido proceso, tomada como parte de la vida ordinaria, podría parecer que requiere una oportunidad para una interacción cara a cara, en la que un ser humano, autorizado a ejercer su discrecionalidad, considere una amplia gama de factores.

En muchas culturas, este argumento en favor del juicio caso por caso tiene una profunda fundamentación moral. Puede encontrarse en la política, en el derecho, en la teología y hasta en la literatura. *El mercader de Venecia*, de Shakespeare, se lee fácilmente como una objeción a las normas sin ruido y un alegato en favor del papel de la clemencia en la ley y en el juicio humano en general. He aquí el argumento final de Porcia:

> La clemencia no es cualidad forzosa.
> Cae como la lluvia, desde el cielo
> a lo que está debajo. Su bendición es doble:
> bendice al que la da y al que la obtiene.
> [...]
> Está entronizada en el alma del rey,
> es atributo del mismo Dios,
> y el poder terrenal más se acerca al de Dios,
> si la clemencia suaviza la justicia.

Al no estar sujeta a normas, la clemencia es ruidosa. No obstante, el alegato de Porcia puede hacerse en muchas situaciones y en innumerables organizaciones. A menudo resuena. Un empleado puede estar buscando un ascenso. Un aspirante a propietario de una vivienda puede estar pidiendo un préstamo. Un estudiante puede estar solicitando una plaza en la universidad. Quienes toman las decisiones en estos casos podrían rechazar algunas estrategias de reducción del ruido, sobre todo las normas estrictamente establecidas, y, si no lo hacen, podría deberse a que piensan, con Porcia, que la clemencia no es algo forzado. Pueden saber que su particular método es ruidoso, pero, si garantiza que la gente sienta que ha sido tratada con respeto y que alguien la ha escuchado, puede que de todos modos lo acepten.[1]

Algunas estrategias de reducción del ruido no se enfrentan a esta objeción. Si tres personas, en lugar de una sola, toman una decisión, se escucha a las personas una a una. Las directrices pueden dejar a quienes deciden con una enorme discrecionalidad. Sin embargo, algunos intentos de reducción del ruido, incluidas las rígidas normas, eliminan esa discrecionalidad y pueden hacer que la gente objete que el proceso resultante ofende a su sentido de la dignidad.

¿Tienen razón? Sin duda, a la gente suele preocuparle que no se la escuche de manera individual. Hay un valor humano incuestionable en la

oportunidad de ser escuchado. Pero, si las audiencias individualizadas producen más muertes, más injusticia y costes mucho más elevados, no deberían celebrarse. Hemos subrayado que, en situaciones como las contrataciones, las admisiones y las propias de la medicina, algunas estrategias de reducción del ruido pueden ser rígidas; pueden prohibir formas de trato individualizado que, aunque sean ruidosas, producen menos errores. Sin embargo, si una estrategia de reducción del ruido es rígida, entonces, como hemos insistido, la mejor respuesta es tratar de encontrar otra mejor, una que se adapte a una amplia gama de variables relevantes. Y, si esa estrategia mejor elimina el ruido y produce menos errores, tendrá ventajas obvias sobre el tratamiento individualizado, aunque reduzca o elimine la oportunidad de ser escuchado.

No estamos diciendo que el interés por el tratamiento individualizado no tenga importancia, pero hay que pagar un alto precio si ese tratamiento conduce a todo tipo de consecuencias terribles, incluida una obvia injusticia.

CAMBIO DE VALORES

Imaginemos que una institución pública consigue eliminar el ruido. Supongamos que una universidad define la «mala conducta» de manera que cada miembro del profesorado y cada estudiante sepa lo que incluye y lo que no. Supongamos que una gran empresa especifica exactamente qué significa «corrupción» para que cualquier persona de la empresa sepa lo que está permitido y lo que está prohibido. O imaginemos que una institución privada reduce el ruido de forma significativa, quizá diciendo que no contratará a nadie que no se haya especializado en determinadas materias. ¿Qué ocurre si los valores de una organización cambian? Algunas estrategias de reducción del ruido parecerían incapaces de hacerles un sitio, y su inflexibilidad podría ser un problema estrechamente relacionado con el interés por el tratamiento individualizado y la dignidad.

Una famosa y desconcertante decisión dentro del derecho constitucional estadounidense nos ayudará a aclarar este punto.[2] El caso, resuelto en 1974, se refería a una norma estricta del sistema escolar que obligaba a las profesoras embarazadas a obtener un permiso no remunerado de cinco meses antes de la fecha prevista para el parto. Jo Carol LaFleur, una

profesora, argumentó que podía continuar muy bien con su labor educativa, que la norma era discriminatoria y que cinco meses eran demasiados.

El Tribunal Supremo de Estados Unidos le dio la razón. Sin embargo, no habló de discriminación por razón de sexo ni dijo que cinco meses fueran forzosamente demasiados. En cambio, objetó que no se había dado a LaFleur la oportunidad de demostrar que no había necesidad física de que ella, en particular, dejara de trabajar. Estas fueron las palabras del tribunal:

> No existe una determinación individualizada por parte del médico de la profesora —o del consejo escolar— en cuanto a la capacidad de cualquier profesora para continuar en su trabajo. Las normas contienen una presunción irrefutable de incompetencia física, y esa presunción se aplica aun cuando las pruebas médicas sobre el estado físico de una mujer concreta puedan ser totalmente contrarias.

Un periodo obligatorio de cinco meses de baja parece absurdo, pero el tribunal no hizo hincapié en este punto. En cambio, lamentó la «presunción irrefutable» y la ausencia de una «resolución individualizada». Por lo visto, al decir esto, el tribunal estaba argumentando, como Porcia, que la calidad de la clemencia no es algo forzoso y que se debe exigir que una persona concreta examine las circunstancias particulares de LaFleur.

Sin embargo, sin higiene alguna de la decisión, esta es una fórmula para el ruido. ¿Quién decide el caso de LaFleur? ¿Será esta decisión la misma para ella que para muchas otras mujeres en una situación similar? En cualquier caso, muchas normas equivalen a presunciones irrefutables. ¿Es inaceptable un límite de velocidad determinado? ¿Una edad mínima para votar o beber? ¿Una prohibición absoluta de conducir bajo los efectos del alcohol? Recurriendo a estos ejemplos, los críticos objetaron que un argumento contra las «presunciones irrefutables» resultaría excesivo (en particular porque el propósito y efecto de estas es reducir el ruido).

Influyentes comentaristas de la época defendieron la decisión del tribunal haciendo hincapié en que los valores morales cambian con el tiempo y, por tanto, en la necesidad de evitar normas rígidas.[3] Argumentaron que, con respecto al papel de la mujer en la sociedad, las normas sociales estaban cambiando de forma acelerada. Sostenían que las resoluciones individualizadas eran muy adecuadas en ese contexto, porque permitirían incorporar esas normas cambiantes. Un sistema de reglas podría eliminar el

ruido, lo cual está bien, pero asimismo podría congelar las normas y valores existentes, lo cual no está tan bien.

En resumen, algunas personas podrían insistir en que una ventaja de un sistema ruidoso es que permitirá a la gente acomodar valores nuevos o incipientes. A medida que los valores cambian, y se permite a los jueces ejercer su discrecionalidad, podrían empezar a dictar, por ejemplo, sentencias más leves a los condenados por delitos relacionados con las drogas, y más severas a los condenados por violación. Hemos subrayado que, si algunos jueces son indulgentes y otros no, habrá algún grado de injusticia; personas en situación similar serán tratadas de forma diferente. Sin embargo, la injusticia podría tolerarse si da cabida a valores sociales nuevos o incipientes.

El problema no se limita al sistema de justicia penal o incluso a la ley. Con respecto a cualquier número de políticas, las empresas podrían decidir permitir cierta flexibilidad en sus juicios y decisiones, aunque hacerlo produzca ruido, porque la flexibilidad garantiza que, a medida que surjan nuevas creencias y valores, las políticas puedan cambiar con el tiempo. Ofreceremos un ejemplo personal: cuando, hace algunos años, uno de nosotros se incorporó a una gran empresa de consultoría, el no tan reciente mensaje de bienvenida que recibió especificaba los gastos de viaje por los que se le permitía reclamar el reembolso («una llamada telefónica a casa tras llegar sin problemas; el gasto de un planchado del traje; propinas para los botones»). Las normas no eran ruidosas, pero estaban claramente anticuadas (y eran sexistas). Pronto fueron sustituidas por normas que podían evolucionar con el tiempo. Por ejemplo, ahora los gastos deben ser «adecuados y razonables».

La primera respuesta a esta defensa del ruido es sencilla: algunas estrategias de reducción del ruido no se enfrentan en absoluto a esta objeción. Si las personas utilizan una escala compartida basada en una visión externa, pueden responder a los cambios de valores a lo largo del tiempo. En cualquier caso, los intentos de reducción del ruido no necesitan ni deben ser permanentes. Si esos intentos adoptan la forma de normas rígidas, quienes las establecen deben estar dispuestos a hacer cambios con el tiempo. Podrían revisarlas anualmente. Podrían decidir que, debido a los nuevos valores, son esenciales nuevas normas. En el sistema de justicia penal, los responsables de las normas pueden reducir las penas para ciertos

delitos y aumentarlas para otros, y pueden despenalizar una actividad y penalizar otra que antes se consideraba totalmente aceptable.

Pero demos un atrás. Los sistemas ruidosos pueden dar cabida a valores morales incipientes, y eso puede ser bueno. Sin embargo, en muchas esferas, es absurdo defender altos niveles de ruido con este argumento. Algunas de las estrategias más importantes para reducir el ruido, como la suma de juicios, aceptan los nuevos valores. Y si varios clientes que se quejan de que un ordenador portátil funciona mal son tratados de forma diferente por una empresa de informática, es poco probable que esta falta de coherencia se deba a valores incipientes. Si distintas personas reciben diferentes diagnósticos médicos, rara vez se debe a nuevos valores morales. Podemos hacer mucho para reducir el ruido, o incluso eliminarlo, y al mismo tiempo diseñar procesos que permitan que los valores evolucionen.

FRAUDE AL SISTEMA: ELUSIÓN DE LAS NORMAS

En un sistema ruidoso, juzgadores de todo tipo pueden adaptarse a los requerimientos de la situación y responder a acontecimientos inesperados. Sin embargo, al eliminar el poder de adaptación, algunas estrategias de reducción del ruido pueden tener la consecuencia no intencionada de proporcionar un incentivo para defraudar al sistema. Un posible argumento para tolerar el ruido es que este puede ser un subproducto de los métodos que instituciones privadas y públicas adoptan para evitar ese tipo de fraude.

El código tributario es un ejemplo conocido. Por un lado, el sistema tributario no debe ser ruidoso. Ha de ser claro y predecible; contribuyentes idénticos no han de recibir un trato diferente. Pero, si eliminásemos el ruido en el sistema fiscal, los contribuyentes inteligentes encontrarían de forma inevitable una manera de sortear las normas. Entre los especialistas en fiscalidad existe un acalorado debate sobre si es mejor tener normas claras que eliminen el ruido o, por el contrario, aceptar un grado de vaguedad que permita la imprevisibilidad, pero que también reduzca el riesgo de que las normas claras den lugar a comportamientos oportunistas o interesados.

Algunas empresas y universidades prohíben que la gente incurra en «irregularidades», sin especificar lo que significa. El resultado inevitable es el ruido, que no es bueno, e incluso puede ser muy nocivo. Sin embargo, si

hay una lista específica de irregularidades, se acabarán tolerando comportamientos nada aceptables que no consten explícitamente en la lista.

Como las normas tienen límites claros, la gente puede eludirlas e incurrir en conductas que técnicamente están exentas, pero que causan los mismos perjuicios u otros similares. (Todos los padres de un hijo adolescente lo saben.) Cuando no podemos diseñar fácilmente normas que prohíban todas las conductas que deberían estar prohibidas, tenemos una razón particular para tolerar el ruido, o eso es lo que dice la objeción.

En algunas circunstancias, las reglas claras y definidas que eliminan el ruido comportan el riesgo de su elusión. Y este riesgo podría ser una razón para adoptar alguna otra estrategia para reducir el ruido, como la suma, y quizá para tolerar un enfoque que permita algo de ruido. Pero las palabras «podrían» ser cruciales. Tenemos que preguntarnos cuánta elusión habría, y cuánto ruido. Si hay poca elusión y mucho ruido, entonces nos convienen los enfoques que reducen el ruido. Volveremos sobre esta cuestión en el capítulo 28.

DISUASIÓN Y AVERSIÓN AL RIESGO

Supongamos que el objetivo es disuadir las conductas reprobables, ya sea de los empleados, de los estudiantes o de los ciudadanos de a pie. Un poco de imprevisibilidad, o incluso mucha, podría no ser lo peor. Un empresario podría pensar: «Si la sanción por ciertos tipos de mal comportamiento es una multa, una suspensión o un despido, entonces mis empleados no incurrirán en malos comportamientos». Quienes operan en un sistema de justicia penal podrían pensar: «No nos importa demasiado que los potenciales delincuentes tengan que adivinar la pena que les pudieran imponer. Si la perspectiva de una lotería penal disuade a la gente de cruzar la línea, tal vez se pueda tolerar el ruido resultante».

En teoría, estos argumentos no pueden desestimarse, pero no son del todo convincentes. A primera vista, lo que importa es el valor esperado de la punición, y una probabilidad del 50 por ciento de una multa de cinco mil dólares es equivalente a la certeza de una multa de dos mil quinientos dólares. Por supuesto, algunas personas pueden poner el foco en el peor de los casos. Las personas con aversión al riesgo podrían ser más disuadidas por el 50 por ciento de posibilidades de una multa de cinco mil dólares,

pero la disuasión será menor entre las que asumen el riesgo. Para saber si un sistema ruidoso es más disuasivo, tenemos que saber si los potenciales infractores tienen aversión al riesgo o lo asumen. Y, si queremos reforzar la disuasión, ¿no sería mejor aumentar la pena y eliminar el ruido? De ese modo se acabaría también con la injusticia.

CREATIVIDAD, MORAL E IDEAS NUEVAS

¿Pueden algunos métodos de reducción del ruido suprimir la motivación y el compromiso? ¿Pueden afectar a la creatividad e impedir que las personas hagan grandes progresos? Muchas organizaciones así lo creen. En algunos casos, puede que tengan razón. Para saber si es así, debemos especificar la estrategia de reducción de ruido a la que se oponen.

Recordemos la reacción sumamente negativa de muchos jueces a las directrices de las sentencias. Como dijo uno de ellos: «Debemos aprender de nuevo a confiar en el ejercicio del juicio en las salas».[4] En general, a las personas investidas de autoridad no les agrada que les priven de su discrecionalidad. Pueden sentirse disminuidas y limitadas, incluso humilladas. Cuando se toman medidas para reducir su discrecionalidad, muchas se rebelan. Valoran la oportunidad de ejercer su juicio; incluso pueden amarla. Si se les priva de su discrecionalidad para que hagan lo que hacen todos, pueden sentirse como engranajes de una máquina.

En resumen, un sistema ruidoso puede ser bueno para la moral no porque sea ruidoso, sino porque permite a las personas decidir como mejor les parezca. Si se permite a los empleados responder a las quejas de los clientes a su manera, evaluar a sus subordinados como mejor les parezca o fijar las primas que consideren oportunas, puede que disfruten más de su trabajo. Si la empresa toma medidas para eliminar el ruido, los empleados podrían pensar que su voluntad está restringida, porque estarán siguiendo reglas en lugar de ejercer su particular creatividad. Su trabajo parecerá más mecánico, incluso robotizado. ¿Quién quiere trabajar en un lugar que anula su personal capacidad para tomar decisiones independientes?

Las organizaciones pueden responder a estos sentimientos no solo porque los honran, sino también porque quieren dar a las personas un margen para que aporten nuevas ideas. Una norma puede reducir el ingenio y la inventiva.

Estos puntos son aplicables a muchas personas en el seno de una organización, aunque, por supuesto, no a todas. Las distintas tareas deben evaluarse de forma diferente; los diagnósticos ruidosos de faringitis estreptocócica o hipertensión podrían no ser un buen lugar para ejercer la creatividad. Sin embargo, podríamos estar dispuestos a tolerar el ruido si con ello conseguimos una plantilla más contenta e inspirada. La desmoralización es de por sí un coste que conduce a otros, como el bajo rendimiento. Sin duda, deberíamos ser capaces de reducir el ruido sin dejar de estar abiertos a nuevas ideas. Algunas estrategias de reducción del ruido, como la estructuración de juicios complejos, hacen exactamente eso. Si queremos reducir el ruido manteniendo una buena moral, podríamos seleccionar estrategias de higiene de las decisiones que tengan esa consecuencia. Y los responsables podrían dejar claro que, aunque existan normas fijas, hay un procedimiento para cuestionarlas y replantearlas, pero no para romperlas ejerciendo la discrecionalidad caso por caso.

En una serie de libros de gran viveza, Philip Howard, un distinguido abogado y pensador, expone argumentos similares a favor de permitir juicios más flexibles.[5] Howard quiere que las políticas no adopten la forma de normas prescriptivas que eliminan el ruido, sino de principios generales: «Sea razonable», «Actúe con prudencia», «No asuma riesgos excesivos».

Según Howard, el mundo moderno, plagado de normas públicas, se ha vuelto loco simplemente por la rigidez que caracteriza a estas. Profesores, agricultores, constructores, enfermeros, médicos... Todos estos expertos, y muchos más, viven agobiados por normas que les dicen lo que tienen que hacer y cómo deben hacerlo exactamente. Howard cree que sería mucho mejor permitir a la gente emplear su propia creatividad para averiguar cómo lograr los objetivos pertinentes, ya se trate de mejores resultados educativos, la reducción de los accidentes, un agua más limpia o pacientes más sanos.

Howard ofrece algunos argumentos sugerentes, pero es importante preguntarse por las consecuencias de los métodos que prefiere, incluido el aumento potencial del ruido y el sesgo. A la mayoría de la gente no le gusta la rigidez en abstracto, pero puede ser la mejor manera de reducir el ruido y eliminar el sesgo y el error. Si solo existen principios generales, el ruido en su interpretación y aplicación será inevitable. Un ruido que podría ser

intolerable, incluso escandaloso. Como mínimo, hay que tener muy en cuenta los costes del ruido, cosa que en general no se hace. Cuando vemos que el ruido produce una injusticia generalizada y elevados costes, a menudo llegamos a la conclusión de que es inaceptable y debemos definir estrategias de reducción del ruido que no comprometan valores importantes.

SOBRE LA DIGNIDAD

- «Las personas valoran y necesitan las interacciones cara a cara. Quieren que un ser humano real escuche sus preocupaciones y quejas, y que sea capaz de mejorar las cosas. Claro que esas interacciones producirán inevitablemente ruido. Sin embargo, la dignidad humana no tiene precio.»
- «Los valores morales evolucionan continuamente. Si lo bloqueamos todo, no dejaremos espacio para los valores cambiantes. Algunos intentos de reducción del ruido son demasiado rígidos; impedirían el cambio moral.»
- «Si se quiere disuadir a las personas de las malas conductas, hay que tolerar algo de ruido. Si los estudiantes temen las posibles sanciones por copiar, nos alegraremos: evitarán copiar. Un poco de incertidumbre en forma de ruido puede aumentar la disuasión.»
- «Si eliminamos el ruido, podemos acabar con las normas claras, que los infractores sabrán eludir. El ruido puede ser un precio que merezca la pena pagar si es una forma de evitar comportamientos estratégicos u oportunistas.»
- «Las personas creativas necesitan un margen para ejercer su creatividad. Las personas no son robots. Sea cual fuere su trabajo, merecen un poco de espacio para maniobrar. Si están constreñidas, puede que no sean ruidosas, pero no se recrearán mucho y no podrán aportar sus ideas originales.»
- «Al final, la mayoría de los intentos de defender el ruido no son convincentes. Podemos respetar la dignidad de las personas, dejar mucho espacio para la evolución moral y permitir la creatividad humana sin tolerar la injusticia y los costes del ruido.»

28
¿Normas o estándares?

Si el objetivo es reducir el ruido o decidir si hacerlo (y en qué grado), es útil distinguir entre dos formas de regular el comportamiento: las normas y los estándares. Organizaciones de todo tipo suelen optar por unas u otras formas, o por una combinación de ellas.

En el sector empresarial, una compañía puede decir que los empleados tienen que cumplir con un horario laboral determinado, que nadie puede tomarse vacaciones de más de dos semanas y que, si alguien filtra información a la prensa, será despedido. Otra posibilidad es decir que los empleados deben trabajar «durante una jornada laboral razonable», que las vacaciones se decidirán «caso por caso, en función de las necesidades de la empresa» y que las filtraciones «se sancionarán adecuadamente».

En derecho, una norma puede decir que nadie puede superar un límite numérico de velocidad, que los trabajadores no pueden estar expuestos a agentes cancerígenos o que todos los medicamentos recetados deben ir acompañados de advertencias específicas. Alternativamente, un estándar puede decir que la gente debe conducir «con prudencia», que los empresarios han de proporcionar lugares de trabajo seguros «en la medida de lo posible» o que, al decidir si ofrecen advertencias para los medicamentos recetados, las compañías deben en este respecto actuar «razonablemente».

Estos ejemplos ilustran la principal distinción entre normas y estándares. Las normas están pensadas para eliminar la discrecionalidad; los estándares están pensados para conceder esa discrecionalidad. Siempre que existan normas, el ruido se reducirá de forma considerable. Quienes interpretan las normas deben responder a una cuestión de hecho: ¿a qué velocidad iba el conductor? ¿Se expuso un trabajador a un agente cancerígeno? ¿Tenía el medicamento las advertencias requeridas?

Cuando existen normas, la tarea de determinar los hechos puede implicar un juicio y, por tanto, producir ruido o estar afectada de un sesgo. Son muchos los ejemplos que nos hemos encontrado. Sin embargo, las personas

que establecen las normas pretenden reducir esos riesgos, y, cuando una norma consiste en un número («Nadie puede votar hasta que no cumpla dieciocho años» o «El límite de velocidad es de ciento veinte kilómetros por hora»), el ruido debería reducirse. Las normas tienen una característica importante: «reducen el papel del juicio». Al menos en este sentido, los juzgadores (entendidos en general como todos los que hacen valer normas) tienen menos trabajo. Se atienen a las normas. Para bien o para mal, tienen mucho menos margen de maniobra.

Los estándares son del todo diferentes. Cuando existen estándares, los juzgadores tienen que trabajar mucho para especificar el significado de unos términos abiertos. Pueden tener que hacer abundantes juicios para decidir lo que puede considerarse (por ejemplo) «razonable» o «factible». Además de encontrar hechos, deben dar contenido a enunciados relativamente vagos. De hecho, quienes definen los estándares dejan en manos de otros la autoridad para tomar decisiones. Delegan el poder.

Los tipos de directrices analizados en el capítulo 22 pueden ser normas o estándares. Si son normas, limitan drásticamente el juicio. Incluso si son estándares, pueden estar lejos de estar abiertos. Las puntuaciones de Apgar son directrices, no normas. No prohíben el ejercicio de la discrecionalidad. Cuando las directrices se hacen más estrictas para eliminar esa discrecionalidad, se convierten en normas. Los algoritmos funcionan como normas, no como estándares.

DIVISIONES E IGNORANCIA

Debe quedar claro desde el principio que, cuando empresas, organizaciones, sociedades o grupos están muy divididos, puede ser mucho más fácil crear estándares que normas. Los directivos de empresas pueden estar de acuerdo en que los gestores no deben actuar de forma abusiva sin saber exactamente qué significa esa prohibición. Los gestores podrían oponerse al acoso sexual en el lugar de trabajo sin decidir si el coqueteo es aceptable. Una universidad podría prohibir a los estudiantes copiarse unos a otros sin especificar el significado exacto de ese término. La gente puede estar de acuerdo en que una constitución debe proteger la libertad de expresión sin decidir si ha de proteger la publicidad comercial, las amenazas o la obscenidad. Y puede estar de acuerdo en que las regulaciones

medioambientales deben dictar reglas prudentes para reducir las emisiones de gases de efecto invernadero sin definir en qué consiste tal prudencia.

Establecer estándares sin especificar los detalles puede dar lugar a ruido, el cual podría controlarse mediante algunas de las estrategias que hemos expuesto, como la agregación de juicios y el uso del protocolo de evaluaciones mediadoras. Es posible que los líderes quieran establecer normas, pero que, en la práctica, no puedan ponerse de acuerdo sobre ellas. Las propias constituciones incluyen muchos estándares (que protegen, por ejemplo, la libertad religiosa). Lo mismo ocurre con la Declaración Universal de los Derechos Humanos («Todos los seres humanos nacen libres e iguales en dignidad y derechos»). La gran dificultad de conseguir que distintas personas se pongan de acuerdo sobre las normas para reducir el ruido es una de las razones por las que se establecen estándares y no normas. Los directivos de una empresa pueden ser incapaces de ponerse de acuerdo sobre las palabras concretas que han de regir el trato de los empleados con los clientes. Establecer estándares puede ser lo mejor que pueden hacer esos directivos. Se dan analogías en el sector público. Los legisladores pueden llegar a un compromiso sobre un estándar (y tolerar el ruido resultante) si ese es el precio de promulgar una determinada ley. En medicina, los doctores pueden llegar a un acuerdo sobre los estándares para el diagnóstico de enfermedades; en cambio, los intentos de elaborar normas pueden originar un desacuerdo irresoluble.

Sin embargo, las divisiones sociales y políticas no son la única razón por la que se recurre a los estándares en lugar de a las normas. A veces, el verdadero problema es que se carece de la información que permitiría establecer normas sensatas. Una universidad puede ser incapaz de establecer normas que rijan sus decisiones sobre la promoción de un miembro del profesorado. Un empleador puede tener dificultades para prever todas las circunstancias que le lleven a mantener o disciplinar a sus empleados. Una legislación nacional podría no saber cuál es el nivel adecuado de contaminantes atmosféricos (partículas, ozono, dióxido de nitrógeno, plomo). Lo mejor que puede hacer es establecer algún estándar y confiar en expertos para especificar su significado, aunque la consecuencia sea el ruido.

Las normas pueden estar sesgadas de muchas maneras. Una norma puede prohibir que las mujeres sean policías, o decir que los irlandeses están

excluidos de algún ámbito. Incluso si crean un gran sesgo, las normas reducirán considerablemente el ruido (si todo el mundo las sigue). Si una norma dice que todos los mayores de veintiún años pueden comprar bebidas alcohólicas y que ningún menor de esa edad puede hacerlo, probablemente habrá poco ruido, al menos mientras la gente siga la norma. No obstante, las normas invitan al ruido.

JEFES CONTROLADORES

La distinción entre normas y estándares tiene una gran importancia para todas las instituciones públicas y privadas, incluidas empresas de todo tipo. La elección entre ambas cosas se produce siempre que un superior intenta controlar a un agente. Como se describe en el capítulo 2, los agentes de seguros se esfuerzan por cobrar la prima Ricitos de Oro (una prima ni demasiado alta ni demasiado baja) para beneficiar a su compañía. ¿Establecerían sus jefes estándares o normas para guiar a estos agentes? Cualquier jefe de una empresa puede dar instrucciones a sus empleados de forma muy específica o más general («Usen el sentido común» o «Utilicen su mejor criterio»). Un médico puede utilizar uno u otro enfoque cuando ofrece instrucciones a un paciente. «Tómese una pastilla cada mañana y cada noche» es una norma; «Tómese una pastilla cuando crea que la necesita» es un estándar.

Hemos observado que una compañía de redes sociales como Facebook se preocupará forzosamente del ruido y de cómo reducirlo. La compañía puede decir a sus empleados que retiren un *post* cuyo contenido infrinja una norma clara (que prohíbe, por ejemplo, los desnudos), o puede decir a sus empleados que se atengan a un estándar (como la prohibición del acoso o de materiales claramente ofensivos). Los Facebook's Community Standards, que se hicieron públicos por primera vez en 2018, son en realidad una fascinante mezcla de normas y estándares, con mucho de ambas cosas. Tras su publicación, se produjeron numerosas quejas por parte de los usuarios de Facebook, que argumentaban que los estándares de la compañía producían excesivo ruido (y, por tanto, incurrían tanto en errores como en injusticias). Una preocupación recurrente era que, dado que muchos miles de revisores de Facebook tenían que emitir juicios, las decisiones podían ser muy variables. A la hora de decidir si retiraban los *posts* que leían, los revisores

tomaban decisiones diferentes sobre lo que estaba permitido y lo que estaba prohibido. Para ver por qué esa variabilidad era inevitable, considérense estos términos de los Facebook's Community Standards en 2020:

> Definimos la incitación al odio como un ataque directo a personas basado en lo que llamamos «características protegidas»: raza, etnia, origen nacional, afiliación religiosa, orientación sexual, casta, sexo, género, identidad de género y enfermedad o discapacidad grave. También proporcionamos algunas protecciones por la condición de inmigrante. Definimos el ataque como un discurso violento o deshumanizado, las afirmaciones de inferioridad o los llamamientos a la exclusión o la segregación.[1]

Al aplicar una definición de este tipo, los revisores serán por fuerza ruidosos. ¿Qué cuenta exactamente como «discurso violento o deshumanizador»? Facebook era consciente de estas cuestiones y, en respuesta a ellas, se movió en la dirección de normas contundentes, precisamente para reducir el ruido. Esas normas se catalogaron en un documento no público llamado «Implementation Standards», que consta de unas doce mil palabras y que obtuvo *The New Yorker*.[2] En los «Community Standards» públicos, el texto que regula el contenido gráfico comenzaba con un estándar: «Eliminamos contenidos que exalten la violencia». (¿Qué es eso, exactamente?) En cambio, los «Implementation Standards» listaban imágenes gráficas y decían explícitamente a los moderadores de contenido qué hacer con esas imágenes. Los ejemplos incluían «seres humanos carbonizados o en llamas» y «la separación de partes del cuerpo no provocadas». Para resumir un asunto complicado, los «Community Standards» se parecen más a los estándares, mientras que los «Implementation Standards» se asemejan más a las normas.

En la misma línea, una compañía aérea puede pedir a sus pilotos que se atengan a normas o a estándares. La cuestión podría ser si hay que volver a la puerta de embarque tras noventa minutos en la pista o cuándo exactamente hay que encender la señal del cinturón de seguridad. Puede que la compañía prefiera las normas porque limitan la discrecionalidad de los pilotos, lo que reduce los errores. Sin embargo, también puede creer que, en algunas circunstancias, los pilotos deben utilizar su mejor criterio. En estas situaciones, los estándares pueden ser mucho mejores que las normas, aunque produzcan algo de ruido.

En todos estos casos, y en muchos más, quienes deciden entre normas y estándares han de centrarse en el problema del ruido, en el problema del sesgo o en ambos. Las empresas, tanto grandes como pequeñas, tienen que tomar esa decisión continuamente. A veces lo hacen de forma intuitiva y con escaso marco de actuación.

Los estándares tienen muchas formas y tamaños. Pueden carecer de contenidos esenciales: «Haga lo que resulte apropiado, según las circunstancias». Pueden estar redactados de forma que se aproximen a las normas, como, por ejemplo, cuando se define específicamente lo que es apropiado con el fin de limitar la discrecionalidad de quienes juzgan. Las normas y los estándares también pueden mezclarse y combinarse. Por ejemplo, una selección de personal puede adoptar una norma («Todos los candidatos deben tener un título universitario») que precede a la aplicación del estándar («Con la mencionada limitación, elija a personas que harán un excelente trabajo»).

Hemos dicho que las normas deben reducir, o incluso eliminar, si es posible, el ruido y que los estándares suelen producir gran cantidad de ruido (a menos que se adopte alguna estrategia de reducción de este). En las organizaciones privadas y públicas, el ruido suele ser producto de la falta de normas. Cuando el ruido es lo bastante fuerte —cuando cualquiera puede ver que personas en situación similar no reciben un trato similar—, suele producirse un movimiento hacia las normas. Como en el caso de las sentencias penales, el movimiento puede convertirse en un clamor. Algún tipo de auditoría del ruido suele preceder a ese clamor.

LA VUELTA A LAS NORMAS

Consideremos una cuestión importante: ¿quién cuenta como discapacitado, de modo que deba tener derecho a las prestaciones económicas reservadas a quienes no pueden trabajar? Si la pregunta se formula así, los jueces tomarán decisiones *ad hoc* que serán ruidosas y, por tanto, injustas. En Estados Unidos, esas decisiones ruidosas e injustas se convirtieron en su día en la norma, y los resultados fueron escandalosos. Al parecer, dos personas idénticas en silla de ruedas, o con depresión grave, o con dolor crónico, eran tratadas de forma diferente. En respuesta, los funcionarios públicos cambiaron a algo mucho más parecido a una norma: una «matriz de

discapacidad». La matriz exige juicios relativamente mecánicos basados en la educación, la ubicación geográfica y las capacidades físicas restantes. El objetivo era que las decisiones fueran menos ruidosas.

La principal discusión sobre el problema, consignada por el profesor de Derecho Jerry L. Mashaw, da un nombre al intento de eliminar los juicios ruidosos: «justicia burocrática».[3] Merece la pena recordar el término. Mashaw celebra la creación de la matriz como justa en lo fundamental, precisamente porque promete eliminar el ruido. Sin embargo, en algunas situaciones, la promesa de la justicia burocrática puede no cumplirse. Siempre que una institución pasa a tomar decisiones basadas en normas, existe el riesgo de que vuelva el ruido.

Supongamos que las normas producen resultados espantosos en determinados casos. Si es así, los jueces podrían ignorar simplemente las normas, pensando que son demasiado duras. Por esta razón, podrían ejercer su discrecionalidad mediante una forma leve de desobediencia civil, que puede ser difícil de vigilar o incluso de detectar. En las empresas privadas, los empleados ignoran normas rígidas que parecen estúpidas. Del mismo modo, los organismos administrativos encargados de proteger la seguridad y la salud públicas pueden negarse sin más a aplicar los estatutos cuando son demasiado estrictos y reglamentarios. En el derecho penal, la «anulación por el jurado» se refiere a situaciones en las que los jurados se niegan a seguir la ley por considerarla insensatamente rígida y severa.

Siempre que una institución pública o privada intente controlar el ruido mediante normas rígidas, debe estar alerta ante la posibilidad de que las normas propicien la discrecionalidad oculta. Con la política de los «three strikes», la respuesta frecuente de los fiscales —evitar acusar de un delito grave a personas que habían sido condenadas dos veces— era muy difícil de controlar e incluso de detectar.

Cuando esto ocurra, habrá ruido, pero nadie lo oirá. Tenemos que controlar nuestras normas para asegurarnos de que funcionen como es debido. Si no lo hacen, la existencia de ruido puede ser una pista, y las normas deben revisarse.

UN MARCO

En las empresas y en la administración, la elección entre normas y estándares suele hacerse de forma intuitiva, pero puede llevarse a cabo de una forma más aplicada. En una primera aproximación, la elección depende de solo dos factores: (1) los costes de las decisiones y (2) los costes de los errores.

Con los estándares, los costes de las decisiones pueden ser muy elevados para todo tipo de juzgadores solo porque tienen que trabajar para darles contenido. El ejercicio del juicio puede ser oneroso. Si se les dice a los médicos que utilicen su mejor criterio, pueden tener que dedicar tiempo a pensar en cada caso (y los juicios pueden ser muy ruidosos). Si a los médicos se les dan directrices claras para decidir si los pacientes tienen faringitis estreptocócica, sus decisiones podrían resultar rápidas y relativamente sencillas. Si el límite de velocidad es de ciento veinte kilómetros por hora, los agentes de policía no tienen que pensar mucho sobre la velocidad que se permite a los conductores, pero, si el estándar es que los conductores no pueden ir «a una velocidad poco razonable», los agentes tienen que pensar mucho más (y la aplicación del estándar será, casi con seguridad, ruidosa). Con las normas, los costes de las decisiones suelen ser mucho menores.

Aun así, el asunto es complicado. Las normas pueden ser sencillas de aplicar una vez establecidas, pero, antes de establecer una norma, «alguien tiene que decidir cuál». Elaborar una norma puede ser difícil. A veces es prohibitivamente costoso. Por ello, los sistemas jurídicos y las empresas privadas suelen utilizar palabras como «razonable», «prudente» y «factible». Y, por ello, términos como estos desempeñan un papel igual de importante en campos como la medicina y la ingeniería.

Los costes de los errores se refieren al número y a la magnitud de estos. Una cuestión omnipresente es la de si los agentes están bien informados y, por tanto, son fiables, y si practican la higiene de las decisiones. Si se dan ambas cosas, entonces un estándar podría funcionar bien y habría poco ruido. Los directores necesitan imponer normas cuando tienen motivos para desconfiar de sus agentes. Si los agentes son incompetentes o acusan sesgos, y si no pueden aplicar de forma viable la higiene de las decisiones, entonces su actuación debe estar sujeta a normas limitativas. Las organizaciones sensatas entienden bien que la cantidad de discrecionalidad

que conceden está estrechamente relacionada con el nivel de confianza que tienen en sus agentes.

Por supuesto, hay un continuo que discurre desde la absoluta confianza hasta la desconfianza total. Un estándar podría dar lugar a numerosos errores por parte de agentes no del todo fiables, pero, si esos errores son menores, podrían ser tolerables. Una norma podría dar lugar a solo unos pocos errores, pero, si son catastróficos, sería preferible un estándar. Deberíamos ser capaces de ver que no hay ninguna razón «general» para pensar que los costes de los errores son mayores con las normas o con los estándares. Si una norma es perfecta, sin duda no producirá errores. Pero las normas rara vez son perfectas.

Supongamos que la ley dice que solo se puede comprar alcohol si se tienen veintiún años o más. El objeto de la ley es proteger a los adolescentes de los distintos riesgos asociados al consumo de alcohol. Así entendida, la ley producirá muchos errores. Algunas personas de veinte, diecinueve, dieciocho o incluso diecisiete años pueden consumir alcohol sin problemas. Otras personas de veintidós, cuarenta y dos o sesenta y dos años no pueden. Un estándar produciría menos errores si pudiéramos encontrar los términos adecuados y la gente pudiera emplear esos términos con acierto. Por supuesto, esto resulta muy difícil, y por eso nos encontramos casi siempre con normas simples, basadas en la edad, para la venta de bebidas alcohólicas.

Este ejemplo plantea una cuestión mucho más amplia. Siempre que haya que tomar numerosas decisiones, es muy posible que haya mucho ruido, y existe un argumento de peso para establecer normas claras. Si los dermatólogos atienden a un gran número de pacientes con sarpullidos y lunares que causan picores, pueden cometer menos errores si sus juicios están limitados por normas sensatas. Sin esas normas, y con estándares abiertos, los costes de las decisiones tienden a ser tremendos. En el caso de las decisiones repetidas, hay verdaderas ventajas en avanzar hacia las normas mecánicas en lugar de hacia los juicios *ad hoc*. Las cargas de ejercer la discrecionalidad son considerables, y los costes del ruido, o la injusticia derivada de él, pueden ser intolerables.

Las organizaciones inteligentes son muy conscientes de las desventajas de ambas formas de regular el comportamiento. Recurren a las normas, o a estándares que se aproximan a las normas, como forma de reducir el ruido

(y el sesgo). Y, para minimizar los costes de los errores, están dispuestas a dedicar por adelantado un tiempo y una atención considerables a garantizar que las normas sean precisas (lo suficiente).

¿PROSCRIBIR EL RUIDO?

En muchas situaciones, el ruido es un escándalo. La gente vive con él, pero no tendría por qué. Una respuesta sencilla es pasar de la discrecionalidad abierta o de un estándar vago a una norma o algo similar. Ahora tenemos una idea de cuándo la respuesta simple es la acertada. Sin embargo, incluso cuando una norma no es viable o no es una buena idea, hemos identificado una selección de estrategias para reducir el ruido.

Todo esto plantea una gran pregunta: ¿debe el sistema legal proscribir el ruido? Sería demasiado sencillo responder que sí, pero la ley debería hacer mucho más de lo que hace ahora para controlar el ruido. He aquí una forma de pensar en el problema. El sociólogo alemán Max Weber lamentaba la «justicia de cadí», por la cual entendía el uso de juicios informales y *ad hoc* no disciplinados por normas generales. En opinión de Weber, la justicia de cadí era intolerablemente casuística; era una violación del imperio de la ley. El juez, decía Weber, «no impartía justicia de acuerdo con normas formales y "sin relación con personas concretas". En gran medida ocurría lo contrario: juzgaba a las personas según sus cualidades concretas y en función de la situación concreta, o según la equidad y lo adecuado del resultado concreto».[4]

Esta modalidad de justicia, según Weber, «no conoce reglas racionales de decisión». No nos costaría imaginar a Weber lamentando el intolerable ruido que acompañaba a la justicia de cadí. Weber celebró el auge de los juicios burocráticos, impuestos de antemano. (Recordemos la idea de la justicia burocrática.) Consideraba que las formas especializadas, profesionales y sujetas a reglas eran la etapa final de la evolución del derecho. Sin embargo, mucho después de la época en que Weber escribía, está claro que la justicia de cadí, o algo similar, sigue siendo omnipresente. La cuestión es qué hacer al respecto.

No llegaríamos a decir que la reducción del ruido debería formar parte de la Declaración Universal de los Derechos Humanos, pero, en algunos casos, el ruido puede contarse como una violación de derechos, y, en general, los

sistemas jurídicos de todo el mundo deberían esforzarse mucho más por controlarlo. Pensemos en las sentencias penales, las sanciones civiles por infracciones, la concesión o denegación de asilo, las oportunidades educativas, los visados, los permisos de construcción y las licencias profesionales. Imaginemos que un gran organismo público contrata a cientos o incluso miles de personas y lo hace a la buena de Dios; habría una cacofonía de ruido. O supongamos que una institución encargada de la custodia de menores trata a los niños de forma muy diferente dependiendo de si uno u otro empleado es asignado al caso. ¿Cómo se puede aceptar que la vida y el futuro de un niño dependan de esa lotería?

En muchos casos, la variabilidad de esas decisiones está claramente inducida por sesgos, incluidos sesgos cognitivos identificables y ciertas formas de discriminación. Cuando sucede, las personas tienden a encontrar la situación intolerable, y se puede invocar la ley como correctivo, exigiendo prácticas nuevas y diferentes. Organizaciones de todo el mundo ven el sesgo como un bribón. Tienen razón. Sin embargo, no ven el ruido de igual manera. Deberían.

En muchos ámbitos, el nivel actual de ruido es demasiado alto. Impone elevados costes y produce una terrible injusticia. Lo que aquí hemos catalogado es la punta del iceberg. La ley debería hacer mucho más para reducir esos costes. Tendría que combatir esa injusticia.

SOBRE NORMAS Y ESTÁNDARES

- «Las normas simplifican la vida y reducen el ruido. Sin embargo, los estándares nos permiten tener presentes las particularidades de las situaciones.»
- «¿Normas o estándares? Preguntémonos primero qué produce más errores. Y, después, qué es más fácil o más complicado de establecer o de aplicar.»
- «A menudo nos atenemos a estándares cuando deberíamos adoptar normas (simplemente porque no prestamos atención al ruido).»
- «La reducción del ruido no debe formar parte de la Declaración Universal de los Derechos Humanos, al menos por ahora. Sin embargo, el ruido puede ser tremendamente injusto. En todo el mundo, los sistemas jurídicos deberían considerar la posibilidad de tomar enérgicas medidas para reducirlo.»

Resumen y conclusión
Tomar el ruido en serio

El ruido es la variabilidad no deseada de los juicios, y hay demasiado. Nuestros principales objetivos han sido explicar aquí por qué es así y ver qué se puede hacer al respecto. Hemos empleado una gran cantidad de material en este libro y, a modo de conclusión, haremos ahora un rápida revisión de los puntos más importantes y ofreceremos una perspectiva más amplia.

JUICIOS

El término «juicio» tal como aquí lo utilizamos no debe confundirse con «pensamiento». Es un concepto mucho más específico: juicio es aquí una forma de medición en la que el instrumento es una mente humana. Al igual que otras mediciones, un juicio asigna una puntuación a un objeto. Esta no tiene por qué ser un número. «El tumor de Mary Johnson es probablemente benigno» es un juicio, como lo son afirmaciones como «La economía nacional es muy inestable», «Fred Williams sería la mejor persona para contratar como nuestro nuevo gerente» y «La prima del seguro para este riesgo debería ser de doce mil dólares». Los juicios integran sin ceremonias distintos retazos de información en una evaluación global. No son cálculos ni siguen reglas exactas. Un profesor utiliza el juicio para calificar una redacción, pero no para puntuar un test de selección múltiple.

Muchas personas se ganan la vida haciendo juicios profesionales, y muchas otras se ven afectadas de forma importante por esos juicios. Entre los «juzgadores» profesionales, como los llamamos aquí, se encuentran entrenadores de fútbol americano, cardiólogos, abogados, ingenieros, ejecutivos de Hollywood, agentes de seguros y muchos más. Los juicios profesionales han ocupado el centro de este libro porque se han estudiado ampliamente y porque su calidad tiene importantes consecuencias para todos nosotros. Creemos que lo que hemos aprendido se aplica también a los juicios que la gente hace sobre determinados aspectos de su vida.

Algunos juicios son «predictivos», y algunos juicios predictivos son verificables; el tiempo nos dirá si eran acertados. Este es, por lo general, el caso de las predicciones de resultados a corto plazo, como los efectos de un medicamento, el curso de una pandemia o los resultados de unas elecciones. Sin embargo, muchos juicios, incluidas las predicciones a largo plazo y las respuestas a preguntas ficticias, no son verificables. La calidad de estos juicios solo puede evaluarse por la calidad del proceso intelectual que los produce. Además, muchos juicios no son predictivos, sino «evaluativos»: la sentencia dictada por un juez o la categoría de un cuadro en una muestra pictórica no pueden compararse fácilmente con un valor objetivamente verdadero.

Sin embargo, resulta sorprendente que las personas que emiten juicios se comporten como si existiera un valor real, con independencia de que exista o no. Piensan y actúan como si existiera una diana invisible a la que apuntar y en la que ni ellas ni las demás deberían fallar en exceso. La expresión «juicio personal» implica tanto la posibilidad de desacuerdo como la expectativa de que este sea limitado. Es decir, las cuestiones de juicio personal se caracterizan por la expectativa de un «desacuerdo limitado». Ocupan un espacio entre las cuestiones de cálculo, en las que no se permite el desacuerdo, y las cuestiones de gusto, en las que hay pocas expectativas de acuerdo, salvo en casos extremos.

LOS ERRORES: SESGO Y RUIDO

Decimos que existe un «sesgo» cuando la mayoría de los errores de un conjunto de juicios van en la misma dirección. El sesgo es el «error medio», como, por ejemplo, cuando un equipo de tiradores da sistemáticamente abajo y a la izquierda de la diana, cuando los ejecutivos son demasiado optimistas sobre las ventas año tras año o cuando una empresa sigue reinvirtiendo dinero en proyectos fallidos que debería descartar.

Suprimir el sesgo de un conjunto de juicios no eliminará todos los errores. Los que quedan cuando se anula el sesgo no son compartidos. Son la divergencia no deseada de los juicios, la falta de fiabilidad del instrumento de medición que aplicamos a la realidad. Son el «ruido». El ruido es la variabilidad en juicios que deberían ser idénticos. Utilizamos el término «ruido del sistema» para el ruido observado en las organizaciones

que emplean profesionales intercambiables para tomar decisiones, como los médicos en una sala de urgencias, los jueces que dictan sentencias y los agentes de una compañía de seguros. Gran parte de este libro se ha ocupado del ruido del sistema.

MEDICIÓN DEL SESGO Y DEL RUIDO

El «error cuadrático medio» (ECM) ha sido el estándar de exactitud en la medición científica durante doscientos años. Las principales características del ECM son que proporciona la media de una muestra como una estimación no sesgada de la media de la población, que trata por igual los errores positivos y negativos y penaliza de forma desproporcionada los errores grandes. El ECM no refleja los costes reales de los errores de juicio, que suelen ser asimétricos. Sin embargo, las decisiones profesionales siempre requieren predicciones exactas. Para una ciudad que se enfrenta a un huracán, está claro que los costes de subestimar y sobrestimar la amenaza no son los mismos, pero no se querrá que estos costes influyan en la predicción de los meteorólogos sobre la velocidad y la trayectoria del huracán. El ECM es el estándar adecuado para hacer esos juicios predictivos cuando el objetivo es la exactitud objetiva.

Medidos por el ECM, el sesgo y el ruido son fuentes de error independientes y aditivas. Obviamente, el sesgo siempre es malo, y reducirlo siempre mejora la exactitud. Menos intuitivo es el hecho de que el ruido es igual de malo y que reducirlo siempre supone una mejora. La mejor cantidad de dispersión es cero, incluso cuando los juicios están claramente sesgados. El objetivo es, por supuesto, minimizar tanto el sesgo como el ruido.

El sesgo en un conjunto de juicios verificables se define por la diferencia entre el juicio medio de un caso y el valor real correspondiente. Esta comparación es imposible en los juicios no verificables. Por ejemplo, nunca se conocerá el valor real de una prima que un asegurador fija para un riesgo concreto. Tampoco podemos conocer fácilmente el valor real de la condena justa para un delito concreto. A falta de ese conocimiento, una suposición frecuente y conveniente (aunque no siempre correcta) es que los juicios no están sesgados y que la media de muchos jueces es la mejor estimación del valor real.

El ruido en un sistema puede evaluarse mediante una «auditoría del ruido», un experimento en el que varios profesionales hacen juicios independientes sobre los mismos casos (reales o ficticios). Podemos medir el ruido sin conocer un valor real, al igual que podemos ver, en el dorso de un blanco, la dispersión de un conjunto de disparos. Las auditorías del ruido pueden medir la variabilidad de los juicios en muchos sistemas, como el servicio de radiología o el sistema de justicia penal. A veces pueden llamar la atención sobre deficiencias en la capacidad o la formación. Y cuantifican el ruido del sistema (por ejemplo, cuando agentes de seguros del mismo equipo difieren en sus evaluaciones de los riesgos).

¿Qué problema es mayor?, ¿el del sesgo o el del ruido? Depende de la situación. La respuesta podría ser el ruido. El sesgo y el ruido contribuyen por igual al error general (ECM) cuando la media de los errores (el sesgo) es igual a las desviaciones estándar de los errores (el ruido). Cuando la distribución de los juicios es normal (la curva estándar en forma de campana), los efectos del sesgo y del ruido son iguales si el 84 por ciento de los juicios está por encima (o por debajo) del valor real. Se trata de un sesgo importante, que a menudo será detectable en un contexto profesional. Cuando el sesgo es inferior a una desviación típica, el ruido es la mayor fuente del error global.

EL RUIDO ES UN PROBLEMA

La variabilidad como tal no es problemática en algunos juicios, incluso es bienvenida. La diversidad de opiniones es esencial para generar ideas y opciones. El pensamiento contrario es esencial para la innovación. La pluralidad de opiniones entre los críticos de cine es algo típico, no una tacha. Los desacuerdos en el comercio crean mercados. Las diferencias de estrategia entre nuevas empresas que compiten entre sí permiten a los mercados seleccionar a la más capaz. Sin embargo, en lo que llamamos cuestiones de juicio, el ruido del sistema es siempre un problema. Si dos médicos hacen diagnósticos diferentes, al menos uno de ellos está equivocado.

Las sorpresas que motivaron este libro son la magnitud de ruido del sistema y la cantidad de daños que causa. Ambas cosas superan con creces las expectativas habituales. Hemos puesto ejemplos de muchos campos,

como el de los negocios, la medicina, la justicia penal, el análisis de huellas dactilares, las predicciones, las calificaciones de personal y la política. De ahí nuestra conclusión: dondequiera que haya juicio, hay ruido, y más del que se piensa.

El gran papel del ruido en el error contradice la creencia común de que los errores aleatorios no importan, porque se «anulan». Esta creencia es errónea. Si hay varios disparos dispersos alrededor del objetivo, no es útil decir que, por término medio, dan en la diana. Si un candidato a un puesto de trabajo obtiene una calificación más alta de la que merece y otro obtiene una más baja, puede contratarse a la persona equivocada. Si una póliza de seguros es excesivamente alta y otra excesivamente baja, ambos errores son costosos para la compañía de seguros; la primera le hace perder clientes, y la segunda le hace perder dinero.

En resumen, podemos estar seguros de que hay error si los juicios varían sin una buena razón. El ruido es perjudicial incluso cuando los juicios no son verificables y el error no puede medirse. Es injusto que se trate de forma diferente a personas en situación similar, y un sistema en el que los juicios profesionales son a todas luces incoherentes pierde credibilidad.

TIPOS DE RUIDO

El ruido del sistema puede desglosarse en «ruido de nivel» y «ruido de patrón». Unos jueces son, en general, más severos y otros más indulgentes; unos analistas son, en general, optimistas y otros pesimistas sobre las perspectivas del mercado; unos médicos recetan más antibióticos que otros. El «ruido de nivel» es la variabilidad de la media de los juicios emitidos por diferentes individuos. La ambigüedad de las escalas de juicio es una de las fuentes de ruido de nivel. Palabras como «probablemente», o números (por ejemplo, «4 en una escala de 0 a 6»), significan cosas diferentes para personas diferentes. El ruido de nivel es una fuente importante de error en los sistemas de juicio y un objetivo importante para las intervenciones destinadas a reducir el ruido.

El ruido del sistema incluye otro componente, generalmente mayor. Con independencia del nivel medio de sus juicios, dos jueces pueden diferir en sus opiniones sobre qué delitos merecen las penas más severas. Sus decisiones a la hora de sentenciar producirán una «clasificación» diferente

de los casos. Llamamos a esta variabilidad «ruido de patrón» (el término técnico es «interacción estadística»).

La principal fuente de ruido de patrón es estable: es la diferencia en las respuestas personales e idiosincrásicas de los jueces ante el mismo caso. Algunas de estas diferencias reflejan principios o valores que los individuos asumen, de forma consciente o no. Por ejemplo, un juez puede ser muy severo con los ladrones de comercios y extraordinariamente indulgente con los que cometen infracciones de tráfico; otro puede mostrar el patrón contrario. Algunos de los principios o valores subyacentes pueden ser bastante complejos, y el juez puede no ser consciente de ellos. Por ejemplo, un juez podría ser relativamente indulgente con los ladrones de cierta edad sin darse cuenta. Por último, una reacción muy personal ante un caso particular también podría ser estable. Una acusada que se parece a la hija del juez bien podría haber evocado el mismo sentimiento de compasión y, por tanto, de indulgencia otro día.

Este «ruido de patrón estable» refleja la singularidad de los jueces: su respuesta a los casos es tan individual como su personalidad. Las sutiles diferencias entre las personas son a menudo atrayentes e interesantes, pero las diferencias se vuelven problemáticas cuando los profesionales operan dentro de un sistema que presupone la coherencia. En los estudios que hemos examinado, el ruido de patrón estable que esas diferencias individuales producen suele ser la mayor fuente de ruido del sistema.

Aun así, las actitudes características de los jueces ante casos concretos no son del todo estables. El ruido de patrón también tiene un componente transitorio, llamado «ruido de ocasión». Detectamos este tipo de ruido si un radiólogo asigna diagnósticos diferentes a la misma imagen en días distintos, o si un dactiloscopista resuelve que dos huellas coinciden en una ocasión, pero no en otra. Como ilustran estos ejemplos, el ruido de ocasión se mide con más facilidad cuando el juzgador no reconoce el caso como visto anteriormente. Otra forma de demostrar el ruido de ocasión es poner de manifiesto el efecto en los juicios de una característica irrelevante del contexto, como cuando los jueces son más indulgentes después de que su equipo de fútbol americano favorito haya ganado, o cuando los médicos recetan más opioides por la tarde.

Los defectos cognitivos de los juzgadores no son la única causa de los errores en los juicios predictivos. La «ignorancia objetiva» suele desempeñar un papel más importante. Algunos hechos son realmente inescrutables, como cuántos nietos tendrá dentro de setenta años un recién nacido o el número de lotería que resultará premiado en un sorteo que se celebrará dentro de un año. Otros pueden conocerse, pero el juzgador no los conoce. La confianza exagerada de la gente en su juicio predictivo subestima su ignorancia objetiva, así como sus sesgos.

Existe un límite en la exactitud de nuestras predicciones, y este es a menudo bastante reducido. Sin embargo, en general nos sentimos cómodos con nuestros juicios. Lo que nos da esta confianza que nos complace es una «señal interna», una recompensa autogenerada por encajar los hechos y el juicio en una historia coherente. La confianza subjetiva en nuestros juicios no está necesariamente relacionada con su acierto objetivo.

La mayoría de la gente se sorprende al saber que la exactitud de sus juicios predictivos no solo es baja, sino también inferior a la de las fórmulas. Incluso los modelos lineales simples construidos a partir de datos limitados, o las reglas sencillas que pueden escribirse en una servilleta, superan sistemáticamente a los juzgadores humanos. La ventaja fundamental de las reglas y los modelos es que se hallan libres de ruido. Tal y como lo experimentamos subjetivamente, el juicio es un proceso sutil y complejo; no tenemos ningún indicio de que la sutileza pueda ser casi siempre ruidosa. Además, nos resulta difícil imaginar que unas reglas sencillas aplicadas de una manera casi automática a menudo sean más exactas que nosotros (pero esto es un hecho sólidamente establecido).

Los «sesgos psicológicos» son, por supuesto, una fuente de error sistemático o sesgo estadístico. Aunque de forma menos clara, también son una fuente de ruido. Cuando los sesgos no son compartidos por todos los juzgadores, cuando están presentes en distintos grados y cuando sus efectos dependen de circunstancias ajenas, los sesgos psicológicos producen ruido. Por ejemplo, si la mitad de los directivos que toman las decisiones de contratación tienen un sesgo en contra de las mujeres y la otra mitad tiene un sesgo a favor de ellas, no habrá un sesgo general, pero el ruido del sistema provocará muchos errores de contratación. Otro ejemplo es el

efecto desproporcionado de las primeras impresiones. Se trata de un sesgo psicológico, pero ese sesgo producirá ruido de ocasión cuando el orden de presentación de las evidencias varíe de forma aleatoria.

Hemos descrito el proceso del juicio como la integración informal de un conjunto de pistas para producir un juicio en una escala. La eliminación del ruido del sistema requeriría, por tanto, que los juzgadores mantuvieran la uniformidad en el uso de las pistas, en las ponderaciones de estas y en el uso de la escala. Incluso dejando aparte los efectos aleatorios del ruido de ocasión, estas condiciones rara vez se cumplen.

El acuerdo suele ser bastante grande en los juicios sobre aspectos individuales. Diferentes contratadores suelen coincidir en sus evaluaciones sobre cuál de dos candidatos es más carismático o más diligente. El proceso intuitivo compartido de «emparejamiento» a través de facetas como la intensidad —como cuando se empareja una nota media alta con una edad de lectura precoz— generalmente producirá juicios similares. Lo mismo ocurre con los juicios basados en un pequeño número de pistas que apuntan siempre en la misma dirección.

Cando un juicio requiere la «ponderación de múltiples pistas en conflicto», aparecen grandes diferencias individuales. Ante el mismo candidato, algunos contratadores darán más peso a las pruebas de brillantez o carisma; otros se dejarán influir más por cuestiones como la diligencia o la calma bajo presión. Cuando las pistas son inconsistentes y no encajan en una historia coherente, diferentes personas darán de forma inevitable más importancia a determinadas pistas e ignorarán otras. El resultado será el ruido de patrón.

LA OSCURIDAD DEL RUIDO

El ruido no es un problema que sobresalga. Rara vez se habla de él y, desde luego, es menos destacado que el sesgo. Probablemente no se haya pensado mucho en él. Dada su importancia, la oscuridad del ruido es un fenómeno interesante de por sí.

Los sesgos cognitivos y otras distorsiones emocionales o motivadas del pensamiento se utilizan a menudo como explicación de los malos juicios. Los analistas invocan el exceso de confianza, el anclaje, la aversión a la pérdida, el sesgo de disponibilidad y otros sesgos para explicar las

decisiones que dieron mal resultado. Estas explicaciones basadas en sesgos son satisfactorias, porque la mente humana ansía explicaciones causales. Siempre que algo va mal, buscamos una causa (y a menudo la encontramos). En muchos casos, la causa parece ser un sesgo.

El sesgo posee una especie de carisma explicativo del que carece el ruido. Si intentamos explicar retrospectivamente por qué una determinada decisión fue errónea, encontraremos fácilmente el sesgo, y jamás el ruido. Solo una «visión estadística» del mundo nos permite ver el ruido, pero esa visión no es natural, pues preferimos las historias causales. La ausencia de pensamiento estadístico en nuestras intuiciones es una de las razones por las que el ruido recibe mucha menos atención que el sesgo.

Otra razón es que los profesionales rara vez ven la necesidad de afrontar el ruido en sus propios juicios y en los de sus colegas. Tras un periodo de formación, los profesionales suelen emitir juicios por su cuenta. Los dactiloscopistas, los aseguradores experimentados y los funcionarios de patentes veteranos rara vez se toman el tiempo de imaginar si sus colegas podrían estar en desacuerdo con ellos (y dedican aún menos tiempo a imaginar si podrían estar en desacuerdo consigo mismos).

La mayoría de las veces, los profesionales confían en su propio juicio. Esperan que sus colegas estén de acuerdo con ellos, y nunca averiguan si realmente lo están. En la mayoría de los campos, es posible que un juicio nunca se evalúe contrastándolo con un valor real, y a lo sumo será sometido a un escrutinio por parte de otro profesional considerado un «experto de respeto». Solo ocasionalmente los profesionales se enfrentarán a un desacuerdo sorprendente, y, cuando eso ocurra, casi siempre encontrarán razones para considerarlo un caso aislado. Las rutinas de las organizaciones también tienden a ignorar o suprimir las pruebas de divergencia entre los expertos de su entorno. Esto es comprensible; desde una perspectiva organizativa, el ruido es una vergüenza.

CÓMO REDUCIR EL RUIDO (Y TAMBIÉN EL SESGO)

Hay razones para creer que algunas personas juzgan mejor que otras. Los mejores juzgadores se caracterizan por su capacidad para realizar tareas específicas, su inteligencia y cierto estilo cognitivo (que se describe mejor como una «mentalidad activamente abierta»). No es de extrañar que los

buenos juzgadores cometan pocos errores graves. Sin embargo, dadas las múltiples fuentes de diferencias individuales, no deberíamos esperar que incluso los mejores juzgadores estén totalmente de acuerdo en problemas de juicio complejos. La ilimitada variedad de circunstancias, personalidades y experiencias que hacen que cada uno de nosotros sea único es también lo que hace al ruido inevitable.

Una estrategia para reducir los errores es la eliminación de los sesgos. Normalmente, las personas intentan suprimir los sesgos de sus juicios corrigiéndolos con posterioridad o dominándolos antes de que puedan afectar a sus juicios. Proponemos una tercera opción, aplicable sobre todo a las decisiones tomadas en grupo: detectar los sesgos en tiempo real designando a un «observador de las decisiones» para que identifique las señales de sesgo (véase el apéndice B).

Nuestra principal sugerencia para reducir el ruido en el juicio es la «higiene de las decisiones». Elegimos este término porque la reducción del ruido es, como en la higiene sanitaria, una prevención contra un enemigo no identificado. Lavarse las manos, por ejemplo, evita que entren en nuestro cuerpo agentes patógenos desconocidos. Del mismo modo, la higiene de la decisión evitará errores sin que sepamos cuáles son. La higiene de las decisiones es tan poco glamurosa como su nombre y, desde luego, menos emocionante que una lucha victoriosa contra sesgos predecibles. Puede que no haya gloria en la prevención de un daño no identificado, pero es muy recomendable.

Un intento de reducción del ruido en una organización debe comenzar siempre con una auditoría del ruido (véase el apéndice A). Una función importante de la auditoría es obtener el compromiso por parte de la organización de tomarse en serio el ruido. Un beneficio esencial es la evaluación de distintos tipos de ruido.

Hemos descrito los éxitos y las limitaciones de los intentos de reducción del ruido en varios ámbitos. Ahora recapitularemos seis principios que definen la higiene de las decisiones, describiremos cómo abordan los mecanismos psicológicos que causan el ruido y mostraremos cómo se relacionan con las técnicas específicas de higiene de las decisiones de las que hemos hablado.

El objetivo del juicio es la exactitud, no la expresión individual. Esta afirmación es nuestra candidata para el primer principio de higiene de las

decisiones en el juicio. Refleja la forma ajustada, específica, en que hemos definido el juicio en este libro. Hemos demostrado que el ruido de patrón estable es un gran componente del ruido del sistema y que este es una consecuencia directa de las diferencias individuales, de la personalidad de los juzgadores, que hacen que distintas personas se formen diferentes opiniones sobre el mismo problema. Esta observación lleva a una conclusión que será tan impopular como ineludible: el juicio no es el lugar para expresar la individualidad.

Para ser claros, los valores personales, la individualidad y la creatividad son necesarios, incluso esenciales, en muchas fases del pensamiento y la toma de decisiones, incluida la elección de objetivos, la formulación de formas innovadoras de abordar un problema y la creación de opciones. Sin embargo, cuando se trata de emitir un juicio sobre estas, las expresiones de la individualidad son una fuente de ruido. Cuando el objetivo es la exactitud y uno espera que otros estén de acuerdo, también debe considerar lo que pensarían otros juzgadores competentes si estuvieran en su lugar.

Una aplicación radical de este principio es la sustitución del juicio por reglas o algoritmos. La evaluación algorítmica garantiza la eliminación del ruido; de hecho, es el único método que puede eliminar el ruido por completo. Los algoritmos ya se utilizan en muchos ámbitos destacados, y cada vez se recurre más a ellos. Pero es improbable que los algoritmos sustituyan al juicio humano en la fase final de las decisiones importantes (y esto lo consideramos algo bueno). Sin embargo, el juicio puede mejorarse, tanto por el uso adecuado de algoritmos como por la adopción de métodos que hagan que las decisiones dependan menos de la idiosincrasia de un profesional. Hemos visto, por ejemplo, cómo las directrices pueden ayudar en las decisiones a limitar la discrecionalidad de los jueces o a promover la homogeneidad en los diagnósticos de los médicos y, por tanto, reducir el ruido y mejorar las decisiones.

Pensar estadísticamente y tener una visión externa del caso. Decimos que un juzgador tiene una visión externa sobre un caso cuando lo considera como miembro de una clase de referencia de casos similares en lugar de como un problema único. Este método se aleja del modo de pensar por defecto, que se centra firmemente en el caso en cuestión y lo integra en una historia causal. Cuando las personas aplican sus experiencias únicas para formarse una idea única del caso, el resultado es el ruido de patrón. La

visión externa es un remedio para este problema; los profesionales que comparten la misma clase de referencia serán menos ruidosos. Además, la visión externa suele aportar valiosas ideas.

El principio de la visión externa favorece el anclaje de las predicciones en estadísticas de casos similares. También conduce a la recomendación de que las predicciones sean moderadas (el término técnico es «regresivas»; véase el apéndice C). La atención a la amplia gama de resultados pasados y a su limitada predictibilidad ayudaría a los responsables de la toma de decisiones a calibrar la confianza en sus juicios. No se puede culpar a las personas de no predecir lo impredecible, pero sí de su falta de humildad predictiva.

Estructurar los juicios en varias tareas independientes. Este principio de «divide y vencerás» lo convierte en necesario el mecanismo psicológico que hemos descrito como «coherencia excesiva», que hace que las personas distorsionen o ignoren la información que no se ajusta a una historia preexistente o incipiente. La exactitud general se resiente cuando las impresiones sobre distintos aspectos de un caso se contaminan entre sí. Para ofrecer una comparación, piénsese en lo que ocurre con el valor probatorio de un conjunto de testigos cuando se les permite comunicarse.

Es posible reducir la coherencia excesiva dividiendo el problema del juicio en una serie de tareas más pequeñas. Esta técnica es similar a la práctica de las entrevistas estructuradas, en las que los entrevistadores evalúan un rasgo cada vez y lo puntúan antes de pasar al siguiente. El principio de la estructuración inspira las directrices de diagnóstico, como la puntuación de Apgar. También está en el corazón del método que hemos denominado «protocolo de evaluaciones mediadoras». Este protocolo descompone un juicio complejo en múltiples valoraciones basadas en hechos, y se propone garantizar que cada una de ellas sea independiente de las demás. Siempre que sea posible, la independencia se protege asignando las evaluaciones a equipos diferentes y minimizando la comunicación entre ellos.

Resistirse a las intuiciones prematuras. Hemos descrito la señal interna de finalización del juicio que da confianza en su juicio a quienes toman decisiones. La falta de voluntad de los responsables de decisiones para renunciar a esta señal gratificante es una razón clave para la resistencia al uso de directrices y algoritmos y otras reglas que les atan las manos. Está

claro que los responsables de las decisiones necesitan sentirse cómodos con su eventual decisión y tener la satisfactoria sensación de confianza en su intuición. Sin embargo, no deben concederse esta recompensa antes de tiempo. Una elección intuitiva informada por una consideración equilibrada y cuidadosa de las pruebas es muy superior a un juicio precipitado. La intuición no tiene por qué estar prohibida, pero debe ser informada, disciplinada y retrasada.

Este principio inspira nuestra recomendación de «secuenciar la información»: los profesionales que emiten juicios no deben recibir información que no necesitan y que podría sesgarlos, incluso si esa información es fidedigna. En la ciencia forense, por ejemplo, es una buena práctica mantener a los dactiloscopistas al margen de otra información sobre un sospechoso. El control del orden del día de una discusión, un elemento clave del protocolo de evaluación, también pertenece a este apartado. Un orden del día eficiente garantizará que los distintos aspectos del problema se consideren por separado y que la formación de un juicio holístico se retrase hasta que el perfil de las evaluaciones no esté completo.

Obtener juicios independientes de múltiples juzgadores, y luego considerar la posibilidad de sumar esos juicios. El requisito de la independencia no se cumple habitualmente en los procedimientos de las organizaciones, sobre todo en las reuniones en las que las opiniones de los participantes están moldeadas por las de los demás. Debido a los efectos en cascada y a la polarización de grupo, las discusiones en el seno de grupo suelen aumentar el ruido. El sencillo procedimiento de recoger los juicios de los participantes «antes» de la discusión revela el alcance del ruido y facilita la resolución constructiva de las diferencias.

Promediar los juicios independientes garantiza la reducción del ruido del sistema (pero no del sesgo). Un solo juicio es una muestra de juicio extraída de la población de todos los juicios posibles, y aumentar el tamaño de la muestra mejora la precisión de las estimaciones. La ventaja de promediar es aún mayor cuando los juzgadores tienen distintas habilidades y patrones de juicio complementarios. La media de un grupo ruidoso puede acabar siendo más exacta que un juicio unánime.

Propiciar los juicios relativos y las escalas relativas. Los juicios relativos son menos ruidosos que los absolutos, porque nuestra capacidad para categorizar objetos en una escala es limitada, mientras que nuestra

capacidad para hacer comparaciones por pares es mucho mejor. Las escalas de juicios que requieren comparaciones serán menos ruidosas que las que requieren juicios absolutos. Por ejemplo, una «escala de casos» requiere que los juzgadores sitúen un caso en una escala definida por ejemplos conocidos por todos.

Los principios de higiene de las decisiones que acabamos de enumerar no solo son aplicables a los juicios recurrentes, sino también a las decisiones importantes únicas, lo que llamamos «decisiones singulares». La existencia de ruido en las decisiones singulares puede parecer ilógica: por definición, no hay variabilidad que medir si se decide una sola vez. Sin embargo, el ruido está ahí, generando errores. El ruido en un equipo de tiradores es invisible si solo vemos al primer tirador en acción, pero la dispersión se haría evidente si viéramos a los demás tiradores. Del mismo modo, la mejor manera de pensar en los juicios singulares es tratarlos como «juicios recurrentes que se hacen una sola vez». Por ello, la higiene de las decisiones también debería mejorarlos.

Aplicar la higiene de las decisiones puede resultar ingrato. El ruido es un enemigo invisible, y una victoria contra un enemigo invisible solo puede ser una victoria invisible. Sin embargo, igual que la higiene corporal, la higiene de las decisiones es vital. Después de una operación realizada con éxito, a cualquiera le gusta creer que es la habilidad del cirujano la que le ha salvado la vida —y así es, por supuesto—, pero, si el cirujano y todo el personal del quirófano no se hubieran lavado las manos, podría haber muerto. Puede que no haya mucha gloria en la higiene, pero hay resultados.

¿CUÁNTO RUIDO?

La batalla contra el ruido no es, por supuesto, la única consideración para quienes toman decisiones y para las organizaciones. Reducir el ruido puede ser demasiado costoso: un instituto podría eliminar el ruido en las calificaciones haciendo que cinco profesores lean todos y cada uno de los trabajos, pero esa carga de trabajo difícilmente se justifica. Un poco de ruido puede ser inevitable en la práctica, un efecto colateral necesario en un sistema en el que el proceso previsto da a cada caso una consideración

individualizada, que no trata a las personas como engranajes de una máquina y que confiere a los responsables de las decisiones una sensación de acto personal. Un poco de ruido puede ser incluso deseable si la variación que crea permite que un sistema se adapte con el tiempo a nuevas realidades (como cuando el ruido refleja valores y objetivos cambiantes y desencadena un debate que conduce a un cambio en la práctica o en la ley).

Quizá lo más importante sea el hecho de que las estrategias de reducción del ruido pueden tener inconvenientes inaceptables. Muchas preocupaciones sobre los algoritmos son exageradas, pero algunas son legítimas. Los algoritmos pueden producir errores estúpidos que un humano nunca cometería, lo que les resta credibilidad, aunque también consigan evitar muchos errores que los humanos cometen. Pueden estar sesgados por un mal diseño o por un entrenamiento con datos inadecuados. Su falta de rostro puede inspirar desconfianza. Las prácticas de higiene de las decisiones también tienen sus inconvenientes: si se gestionan mal, corren el riesgo de burocratizar las decisiones y de desmoralizar a los profesionales, que sienten que se está socavando su autonomía.

Todos estos riesgos y limitaciones merecen una plena consideración, pero que una objeción a la reducción del ruido tenga sentido depende de la estrategia concreta de reducción del ruido que se esté debatiendo. Una objeción a la agregación de juicios —tal vez por ser demasiado costosa— puede no afectar al uso de directrices. Sin duda, siempre que los costes de la reducción del ruido superen los beneficios, no debería llevarse a cabo. Una vez realizado el cálculo de costes y beneficios, este puede revelar un nivel óptimo de ruido que no sea cero. El problema es que, en ausencia de auditorías del ruido, la gente no es consciente de la cantidad de ruido que hay en sus juicios. Cuando este es el caso, invocar la dificultad de reducir el ruido no es más que una excusa para no medirlo.

Los sesgos conducen a errores e injusticias. Y también el ruido. Sin embargo, hacemos mucho menos al respecto. El error de juicio puede parecer más tolerable cuando es aleatorio que cuando lo atribuimos a una causa, pero no es menos perjudicial. Si queremos tomar mejores decisiones sobre asuntos que importan, debemos tomarnos en serio la reducción del ruido.

Epílogo
Un mundo menos ruidoso

Imagínese cómo serían las organizaciones si se rediseñaran para reducir el ruido. Los hospitales, las mesas de contratación, los analistas del futuro de la economía, los organismos públicos, las compañías de seguros, las autoridades de salud pública, los sistemas de justicia penal, los bufetes de abogados y las universidades estarían muy atentos al problema del ruido y se esforzarían por reducirlo. Las auditorías del ruido serían rutinarias; podrían llevarse a cabo cada año.

Los dirigentes de las organizaciones utilizarían hoy algoritmos para sustituir el juicio humano o para complementar su uso en muchas más áreas. Las personas desglosarían los juicios complejos en evaluaciones mediadoras más sencillas. Conocerían la higiene de las decisiones y seguirían sus prescripciones. Se obtendrían y sumarían juicios independientes. Las reuniones tendrían un aspecto muy diferente; los debates estarían más estructurados. La visión externa se integraría de forma más sistemática en el proceso de decisión. Los desacuerdos manifiestos serían más frecuentes y se resolverían de forma más constructiva.

El resultado sería un mundo menos ruidoso. Se ahorrarían grandes cantidades de dinero, se mejorarían la seguridad y la salud públicas, aumentaría la equidad y se evitarían muchos errores evitables. Nuestro objetivo al escribir este libro ha sido llamar la atención sobre esta oportunidad. Esperamos que los lectores se cuenten entre los que la aprovechen.

Agradecimientos

Son muchas las personas a las que debemos estar agradecidos. Linnea Gandhi ha sido la que podríamos llamar nuestra jefa de personal. Ella nos ha ofrecido orientación y ayuda, nos ha mantenido organizados y nos ha hecho sonreír y reír; básicamente ha dirigido el *show*. Aparte de todo esto, nos ha ofrecido numerosas y valiosas sugerencias sobre el manuscrito. No podríamos haberlo organizado sin ella. Dan Lovallo desempeñó un papel importante como coautor de uno de los artículos que dieron pie a este libro. John Brockman, nuestro agente, se mostró entusiasta, esperanzado, agudo y sabio en cada etapa. Le estamos muy agradecidos. Tracy Behar, nuestra principal editora y guía, mejoró el libro en aspectos notables y menudos. Arabella Pike e Ian Straus también aportaron magníficas sugerencias editoriales.

Especial agradecimiento merecen también Oren Bar-Gill, Maya Bar-Hillel, Max Bazerman, Tom Blaser, David Budescu, Jeremy Clifton, Anselm Dannecker, Vera Delaney, Itiel Dror, Angela Duckworth, Annie Duke, Dan Gilbert, Adam Grant, Anupam Jena, Louis Kaplow, Gary Klein, Jon Kleinberg, Nathan Kuncel, Kelly Leonard, Daniel Levin, Sara McLanahan, Barbara Mellers, Josh Miller, Sendhil Mullainathan, Scott Page, Eric Posner, Lucia Reisch, Matthew Salganik, Eldar Shafir, Tali Sharot, Philip Tetlock, Richard Thaler, Barbara Tversky, Peter Ubel, Crystal Wang, Duncan Watts y Caroline Webb, que leyeron y comentaron los borradores de los capítulos y, en algunos casos, el borrador del texto completo. Les agradecemos su generosidad y ayuda.

Tuvimos la suerte de contar con el consejo de muchos grandes investigadores. Julian Parris nos ofreció una ayuda inestimable en muchas cuestiones estadísticas. Nuestros capítulos sobre los logros del aprendizaje automático no habrían sido posibles sin Sendhil Mullainathan, Jon Kleinberg, Jens Ludwig, Gregory Stoddard y Hye Chang. Y nuestra discusión sobre la coherencia del juicio debe mucho a Alex Todorov y a sus colegas de Princeton Joel Martinez, Brandon Labbree y Stefan Uddenberg, así como a Scott Highhouse y Alison Broadfoot. Estos increíbles equipos de

investigadores no solo compartieron solícitamente sus conocimientos, sino que tuvieron el gesto de realizar análisis especiales para nosotros. Por supuesto, cualquier malentendido o error es responsabilidad nuestra. Damos también las gracias a Laszlo Bock, Bo Cowgill, Jason Dana, Dan Goldstein, Harold Goldstein, Brian Hoffman, Alan Krueger, Michael Mauboussin, Emily Putnam-Horstein, Charles Scherbaum, Anne-Laure Sellier y Yuichi Shoda por compartir sus conocimientos con nosotros. También estamos agradecidos a un auténtico ejército de investigadores por la labor que han venido realizando durante años, compuesto entre otros por Shreya Bhardwaj, Josie Fisher, Rohit Goyal, Nicole Grabel, Andrew Heinrich, Meghann Johnson, Sophie Mehta, Eli Nachmany, William Ryan, Evelyn Shu, Matt Summers y Noam Ziv-Crispel. Muchas de las discusiones aquí presentes conectan con áreas fundamentales en las que no tenemos experiencia y, debido a su excelente trabajo, el libro tiene menos sesgos y menos ruido de los que tendría sin ellos.

Por último, el trabajo en un equipo de tres autores y dos continentes ya fue un reto en las temporadas mejores, y el año 2020 no fue la mejor de ellas. No habríamos concluido este libro sin la magia tecnológica de Dropbox y Zoom. Estamos muy agradecidos a las personas que están detrás de estos grandes servicios.

Apéndice A
Cómo llevar a cabo una auditoría del ruido

Este apéndice ofrece una guía práctica para llevar a cabo una auditoría del ruido. Debe leerse desde la perspectiva de un consultor al que una organización ha encargado examinar la calidad de los juicios profesionales que producen sus empleados y efectuar una auditoría del ruido en una de sus unidades.

Como su nombre indica, el objetivo de la auditoría es averiguar la frecuencia del ruido. Sin embargo, una auditoría bien conducida proporcionará información valiosa sobre los sesgos, los puntos ciegos y las deficiencias específicas en la formación de los empleados y en la supervisión de su trabajo. Una auditoría bien realizada ha de recomendar cambios en las operaciones de la unidad, incluso en la doctrina que guía los juicios de los profesionales, la formación que reciben, las herramientas que utilizan para apoyar sus juicios y la supervisión rutinaria de su trabajo. Si se considera que el esfuerzo ha dado frutos, puede extenderse a otras unidades de la organización.

Una auditoría del ruido requiere una cantidad de trabajo considerable y mucha atención a los detalles, porque su credibilidad será con seguridad cuestionada si sus conclusiones revelan fallos significativos. Por tanto, cada detalle de los casos y del procedimiento debe considerarse bajo una mirada escudriñadora. El proceso que describimos tiene por objeto reducir la oposición haciendo que los profesionales, que son los críticos potenciales más importantes de la auditoría, sean sus autores.

Junto con el consultor (que puede ser externo o interno), el elenco de actores pertinente incluye a estos:

• *El equipo del proyecto*. El equipo del proyecto será responsable de todas las fases del estudio. Si los consultores son internos, formarán el núcleo del equipo del proyecto. Si los consultores son externos, un equipo de proyecto interno colaborará estrechamente con ellos. Esto garantizará que las personas de la empresa vean la auditoría como «su» proyecto y que atribuyan a los consultores una función de apoyo. Además de los consultores, que administran la recogida de datos, analizan los resultados y preparan un informe final, el equipo del proyecto debe incluir expertos en la

materia que puedan construir los casos que los juzgadores evaluarán. Todos los miembros del equipo del proyecto han de gozar de una alta credibilidad profesional.

- *Clientes*. Una auditoría del ruido solo será útil si conduce a cambios significativos, lo que requiere la pronta implicación de la dirección de la organización, que es el «cliente» del proyecto. Es de esperar que los clientes se muestren en un principio escépticos sobre la frecuencia del ruido. Este escepticismo inicial es, en realidad, una ventaja si va acompañado de una actitud abierta, de curiosidad por los resultados de la auditoría y de un compromiso para remediar la situación si se confirman las expectativas pesimistas del consultor.
- *Juzgadores*. Los clientes designarán una o varias unidades para ser auditadas. La unidad seleccionada debe estar formada por un número considerable de «juzgadores», profesionales que hacen juicios y toman decisiones similares en nombre de la empresa. Los juzgadores han de ser intercambiables; es decir, si una persona no está disponible para ocuparse de un caso, se asignará el caso a otra y se esperará que emita un juicio similar. Los ejemplos que se introdujeron en este libro fueron las sentencias de los jueces federales y la fijación de las primas de riesgo más las reservas para reclamaciones en una compañía de seguros. Para una auditoría del ruido, lo mejor es seleccionar una tarea de juicio que (1) pueda completarse sobre la base de información escrita y (2) se exprese numéricamente (por ejemplo, en dólares, probabilidades o calificaciones).
- *Director del proyecto*. Un directivo de alto nivel entre el personal administrativo, que debe ser designado como director del proyecto. No se requiere una experiencia profesional específica para esa tarea. Sin embargo, un alto cargo en la organización tiene una importancia práctica para superar los obstáculos administrativos, y es también una demostración de la importancia que la empresa concede al proyecto. La tarea del director del proyecto es proporcionar apoyo administrativo para facilitar todas las fases del proyecto, incluida la preparación del informe final y la comunicación de sus conclusiones a la dirección de la empresa.

ELABORACIÓN DE LOS MATERIALES DEL CASO

Los expertos en la materia que formen parte del equipo del proyecto deberán tener una experiencia reconocida en las tareas de la unidad (por ejemplo, la fijación de primas para riesgos o la evaluación del potencial de posibles inversiones). Ellos serán los encargados de elaborar los casos que se utilizarán en la auditoría. Diseñar una simulación creíble de los juicios que hacen los profesionales en el trabajo es una tarea delicada, sobre todo teniendo en cuenta el examen detallado al que se someterá el estudio si revela serios problemas. El equipo ha de considerar esta cuestión: si los resultados de nuestra simulación indican un alto nivel de ruido, ¿aceptarán las personas de la empresa que hay ruido en los juicios reales de la unidad? Solo merece la pena llevar a cabo la auditoría del ruido si la respuesta es un claro «sí».

Hay más de una manera de conseguir una respuesta afirmativa. La auditoría del ruido de las sentencias descrita en el capítulo 1 resumió cada

caso mediante una breve lista esquemática de atributos relevantes, y obtuvo evaluaciones de dieciséis casos en noventa minutos. La auditoría del ruido en la compañía de seguros descrita en el capítulo 2 utilizó resúmenes detallados y realistas de casos complejos. Los hallazgos de alto nivel de ruido en ambos casos proporcionaron pruebas aceptables del argumento de que, si se encontró mucho desacuerdo en los casos simplificados, el ruido solo podía ser mayor en los casos reales.

Debe prepararse un cuestionario para cada caso con el fin de entender mejor el razonamiento que llevó a cada juzgador a hacer un juicio sobre ese caso. El cuestionario solo tiene que emplearse después de reunir todos los casos. Y debe incluir:

- Preguntas abiertas sobre los factores clave que llevaron al participante a su respuesta.
- Una lista de los hechos del caso que permita al participante calificar su importancia.
- Preguntas que pidan una «visión externa» de la categoría a la que pertenece el caso. Por ejemplo, si los casos exigen valoraciones en dólares, los participantes han de proporcionar una estimación de en qué medida está el caso por debajo o por encima de la media en comparación con todas las valoraciones de casos de la misma categoría.

REUNIÓN CON LOS EJECUTIVOS PREVIA AL LANZAMIENTO DEL PROYECTO

Cuando se disponga de los materiales del caso que se utilizarán en la auditoría, debe programarse una reunión en la que el equipo del proyecto presentará la auditoría a los dirigentes de la empresa. En esa reunión, las discusiones han de considerar los posibles resultados del estudio, incluido el hallazgo de un ruido inaceptable en el sistema. El objetivo de la reunión es escuchar las objeciones al estudio planeado y obtener de la dirección el compromiso de aceptar sus resultados, sean estos los que fueren; sin ese compromiso, carece de sentido pasar a la siguiente fase. Si se plantean objeciones serias, se puede exigir al equipo del proyecto que mejore los materiales del caso y que lo intente de nuevo.

Una vez que los ejecutivos hayan aceptado el diseño de la auditoría del ruido, el equipo del proyecto deberá pedirles que expongan sus expectativas sobre los resultados del estudio. Tendrán que discutir cuestiones como las siguientes:

- «¿Qué nivel de desacuerdo se espera entre un par de respuestas seleccionadas al azar para cada caso?»

- «¿Cuál es el nivel máximo de desacuerdo que sería aceptable desde una perspectiva empresarial?»
- «¿Cuál es el coste estimado de equivocarse en una evaluación en cualquier dirección (demasiado alta o baja) por una cantidad determinada (por ejemplo, el 15 por ciento)?»

Las respuestas a estas preguntas deben documentarse para asegurarse de que se recuerden y se crean cuando lleguen los resultados reales de la auditoría.

ADMINISTRACIÓN DEL ESTUDIO

Los directivos de la unidad auditada deben ser, desde el principio, informados en términos generales de que su unidad ha sido seleccionada para un estudio especial. Sin embargo, es importante que no se utilice el término «auditoría del ruido» para describir el proyecto. Han de evitarse las palabras «ruido» y «ruidoso», sobre todo en descripciones de personas. En su lugar debe utilizarse un término neutro como «estudio sobre la toma de decisiones».

Los directivos de la unidad serán los encargados de la recogida de datos y los responsables de informar a los participantes sobre la tarea, con la participación del director del proyecto y los miembros del equipo del proyecto. La intención del ejercicio debe describirse a los participantes en términos generales, como por ejemplo: «La organización está interesada en saber cómo (los responsables de la toma de decisiones) llegan a sus conclusiones».

Es esencial asegurar a los profesionales que participan en el estudio que las respuestas individuales no serán conocidas por nadie en la organización, incluido el equipo del proyecto. Si es necesario, se podrá contratar a una empresa externa para anonimizar los datos. También es importante subrayar que no habrá consecuencias específicas para la unidad, que fue seleccionada solo como representante de las unidades que realizan tareas de juicio en nombre de la organización. Para garantizar la credibilidad de los resultados, todos los profesionales cualificados de la unidad deberán participar en el estudio. La asignación de media jornada laboral al ejercicio ayudará a convencer a los participantes de su importancia.

Todos los participantes tendrán que realizar el ejercicio al mismo tiempo, pero deberán mantenerse físicamente separados, y se les pedirá que no se

comuniquen mientras el estudio esté en marcha. El equipo del proyecto estará a su disposición para responder a preguntas durante el estudio.

ANÁLISIS Y CONCLUSIONES

El equipo del proyecto se encargará de los análisis estadísticos de los múltiples casos evaluados por cada participante, entre los que se incluirá la medición de la cantidad global de ruido y sus componentes, el ruido de nivel y el ruido de patrón. Si los materiales del caso lo permiten, también identificará los sesgos estadísticos en las respuestas. El equipo del proyecto tendrá la tarea igualmente importante de intentar comprender las fuentes de variabilidad en los juicios examinando las respuestas al cuestionario en las que los participantes explicaron su razonamiento e identificaron los hechos que más influyeron en sus decisiones. Centrándose principalmente en las respuestas extremas de ambos lados de la distribución, el equipo buscará patrones en los datos. Tratará de encontrar indicios de posibles deficiencias en la formación de los empleados, en los procedimientos de la organización y en la información que esta proporciona a sus empleados.

El consultor y el equipo interno del proyecto trabajarán juntos para desarrollar herramientas y procedimientos que apliquen los principios de la higiene de las decisiones y la supresión de sesgos para mejorar los juicios y las decisiones tomadas en la unidad. Es probable que esta etapa del proceso se prolongue durante varios meses. Paralelamente, el consultor y el equipo profesional también elaborarán un informe sobre el proyecto, que presentarán a la dirección de la organización.

Llegados a este punto, la organización habrá realizado una auditoría del ruido de una de sus unidades. Si el esfuerzo se considera un éxito, el equipo ejecutivo podrá decidir realizar un intento más amplio para evaluar y mejorar la calidad de los juicios y las decisiones que se producen en la organización.

Apéndice B
Una lista de control para un observador de las decisiones

Este apéndice presenta un ejemplo genérico de lista de control para ser utilizada por un observador de las decisiones (véase el capítulo 19). En la lista de control que aquí se presenta aparece de forma aproximada la secuencia cronológica de la discusión que conduce a una decisión importante.

Las preguntas sugeridas que siguen a cada punto de la lista de control aportan aclaraciones añadidas. Los observadores de las decisiones deben hacerse estas preguntas mientras observan el proceso de decisión.

Esta lista de control no está pensada para ser utilizada tal cual. Esperamos más bien que sirva de inspiración y de punto de partida para los observadores de las decisiones, que diseñarán su propia lista de control a medida para la observación de sesgos.

1. ENFOQUE DEL JUICIO

1a. Sustitución

__«¿Indican la elección de las pruebas por parte del grupo y el foco de su discusión la sustitución de una pregunta difícil que se le asignó por otra más fácil?»

__«¿Descuidó el grupo un factor importante (o pareció dar peso a uno irrelevante)?»

1b. Visión interna

__«¿Adoptó el grupo la visión externa para parte de sus deliberaciones y trató seriamente de aplicar un juicio comparativo en lugar de un juicio absoluto?»

1c. Diversidad de opiniones

__«¿Hay alguna razón para sospechar que miembros del grupo comparten sesgos, lo que podría llevar a que sus errores estén

correlacionados? Y, a la inversa, ¿se puede pensar en algún punto de vista o experiencia relevantes que no estén representados en este grupo?»

2. PREJUICIOS Y CIERRE PREMATURO

2a. Prejuicios iniciales

__«¿Ganan (alguno de) los responsables de decisiones más con una conclusión que con otra?»

__«¿Alguien se comprometió ya con una conclusión? ¿Hay alguna razón para sospechar de un prejuicio?»

__«¿Expresaron los que no estaban de acuerdo sus opiniones?»

__«¿Existe el riesgo de que se intensifique el compromiso con una línea de acción que no conduzca a nada?»

2b. Cierre prematuro; coherencia excesiva

__«¿Hubo un sesgo accidental en la elección de las consideraciones que se discutieron primeramente?»

__«¿Se consideraron del todo las alternativas y se buscaron activamente las pruebas que las apoyaran?»

__«¿Se suprimieron datos u opiniones incómodos o desatendidos?»

3. PROCESAMIENTO DE LA INFORMACIÓN

3a. Disponibilidad y relevancia excesiva

__«¿Exageran los participantes la relevancia de un acontecimiento por su actualidad, su radicalidad o su relevancia personal sin tener valor diagnóstico?»

3b. Falta de atención a la calidad de la información

__«¿Se basó el juicio en gran medida en anécdotas, historias o comparaciones? ¿Las confirmaban los datos?»

3c. Anclaje

__«¿Tuvieron números de dudosa exactitud o relevancia un papel importante en el juicio final?»

3d. Predicción no regresiva

__«¿Hicieron los participantes extrapolaciones, estimaciones o predicciones no regresivas?»

4. DECISIÓN

4a. Falacia de la planificación

__«Cuando se utilizaron pronósticos, ¿se cuestionaron sus fuentes y su validez? ¿Se utilizó la visión externa para poner en cuestión los pronósticos?»

__«¿Se utilizaron intervalos de confianza para números dudosos? ¿Son suficientemente amplios?»

4b. Aversión a las pérdidas

__«¿Está la propensión al riesgo de los responsables de las decisiones alineada con la de la organización? ¿Es el equipo de decisión excesivamente cauto?»

4c. Sesgo del presente

__«¿Reflejan los cálculos (incluida la tasa de descuento utilizada) el equilibrio de la organización entre prioridades a corto y a largo plazo?»

Apéndice C
Corrección de las predicciones

Las «predicciones emparejadas» son errores causados por nuestra confianza en el proceso de emparejamiento intuitivo (véase el capítulo 14). Hacemos predicciones emparejadas cuando nos basamos en la información que tenemos para hacer un pronóstico y nos comportamos como si esa información fuera completamente (o muy altamente) predictiva del resultado.

Recordemos el ejemplo de Julie, que podía «leer con fluidez cuando contaba cuatro años». La pregunta era:¿cuál es su nota media? Si el lector predijo un 3,8 sobre 4 de nota media universitaria para Julie, juzgó intuitivamente que la Julie de cuatro años estaba en el 10 por ciento superior de su grupo de edad por su edad de lectura (aunque no en el 3-5 por ciento superior). A continuación, supuso de manera implícita que Julie también se situaría en torno al percentil .90 de su clase en términos de nota media. Esto corresponde a una nota media de 3,7 o 3,8 (de ahí la popularidad de estas respuestas).

Lo que hace que este razonamiento sea estadísticamente incorrecto es que sobrestima en exceso el valor diagnóstico de la información disponible sobre Julie. Uno niño precoz de cuatro años no siempre acaba sobresaliendo en el ámbito académico (y, por fortuna, un niño que en un principio tiene dificultades con la lectura no languidecerá para siempre en el último puesto la clase).

De hecho, lo más frecuente es que el rendimiento sobresaliente se vuelva menos sobresaliente. Y, a la inversa, el rendimiento muy pobre mejorará. Es fácil imaginar razones sociales, psicológicas o incluso políticas para esta observación, pero no son necesarias. El fenómeno es puramente estadístico. Las observaciones extremas en una u otra dirección tenderán a ser menos extremas solo porque el rendimiento pasado no está correlacionado del todo con el rendimiento futuro. Esta tendencia se denomina «regresión a la media» (de ahí el término técnico «no regresivo» para las predicciones emparejadas, que no tienen esto en cuenta).

Para expresarlo de forma cuantitativa, el juicio que el lector hizo sobre Julie sería correcto si la edad de lectura fuera un predictor ideal de la nota media universitaria, es decir, si hubiera una correlación de 1 entre los dos factores. Obviamente, no es este el caso.

Hay una forma estadística de hacer un juicio que probablemente sea más correcta. No es intuitiva y es difícil de encontrar, incluso para personas con cierta formación estadística. He aquí el procedimiento. La figura 19 lo ilustra con el ejemplo de Julie.

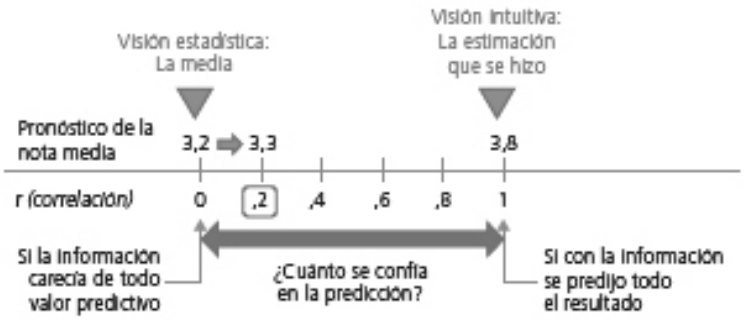

FIGURA 19. Ajuste de una predicción intuitiva por la regresión a la media

1. Haga una estimación intuitiva

La intuición sobre Julie, o sobre cualquier caso sobre el que se tenga información, no es inútil. El pensamiento rápido del Sistema 1 sitúa con facilidad la información que se tiene en la escala de predicción y produce una puntuación para la nota media de Julie. Esta estimación es la predicción que se haría si la información que se tiene fuese completamente predictiva. Anótese.

2. Busque la media

Ahora, se dará un paso atrás para olvidar por un momento lo que se sabe de Julie. ¿Qué diríamos sobre la nota media de Julie «si no supiéramos absolutamente nada sobre ella»? La respuesta es muy sencilla: a falta de cualquier información, la mejor estimación de la nota media de Julie tendría que ser la nota media del último curso (probablemente alrededor de 3,2).

Estimar la nota media de Julie de esta manera es una aplicación del principio más amplio que hemos discutido anteriormente: la «visión

externa». Cuando adoptamos la visión externa, pensamos en el caso que estamos considerando como un ejemplo de una clase, y pensamos en esa clase en términos estadísticos. Recordemos, por ejemplo, cómo el hecho de adoptar la visión externa sobre el problema de Gambardi nos lleva a preguntar cuál es la tasa-base de éxito de un nuevo director ejecutivo (CEO) (véase el capítulo 4).

3. Estime el valor diagnóstico de la información que se tiene

Este es el paso difícil, en el que hay que preguntarse: «¿Cuál es el valor diagnóstico de la información que se tiene?». La razón por la que esta pregunta es importante ya debería estar clara. Si todo lo que supiéramos sobre Julie es su talla de zapatos, daríamos correctamente un peso cero a esta información, y nos ceñiríamos a la predicción de la nota media. Si, por otra parte, tuviéramos la lista de calificaciones que Julie ha obtenido en cada asignatura, esta información sería totalmente predictiva de su nota media (que es su promedio). Hay muchos tonos de gris entre estos dos extremos. Si tuviéramos datos sobre los excepcionales logros intelectuales de Julie en el instituto, esta información tendría mucho más valor diagnóstico que la edad en que aprendió a leer, pero menos que sus calificaciones universitarias.

Para hacer una estimación sensata, recordemos algunos de los ejemplos que enumeramos en el capítulo 12. En las ciencias sociales, las correlaciones de más de .50 son muy raras. Muchas correlaciones que reconocemos como significativas están en el rango de .20. En el caso de Julie, una correlación de .20 sería probablemente un límite superior.

4. Haga un ajuste desde la visión externa, hacia la estimación intuitiva, hasta un punto que refleje el valor diagnóstico de la información que se tiene

El último paso es una simple combinación aritmética de los tres números que ahora se tienen: debe hacerse un ajuste desde la media, hacia la estimación intuitiva, en proporción a la correlación que se ha estimado.

Este paso simplemente amplía la observación que acabamos de hacer: si la correlación fuera 0, nos quedaríamos en la media; si fuera 1, no

tendríamos en cuenta la media y haríamos alegremente una predicción emparejada. En el caso de Julie, por tanto, la mejor predicción que puede hacerse de la media es una que no se aleje más del 20 por ciento de la media de la clase hacia la estimación intuitiva que sugirió su edad de lectura. Este cálculo lleva a una predicción de 3,3, aproximadamente.

Hemos utilizado el ejemplo de Julie, pero este método puede aplicarse con la misma facilidad a muchos de los problemas de juicio que hemos tratado en este libro. Consideremos, por ejemplo, a un vicepresidente de ventas que ha de contratar a un nuevo vendedor y acaba de tener una entrevista con un candidato de todo punto excepcional. Basándose en esta fuerte impresión, el ejecutivo estima que el candidato registraría ventas por valor de un millón de dólares en el primer año de trabajo (el doble de la media alcanzada por los nuevos contratados durante ese periodo). ¿Cómo puede el vicepresidente hacer esta estimación regresiva? El cálculo depende del valor diagnóstico de la entrevista. ¿En qué medida la entrevista de contratación predice el éxito en el puesto de trabajo en este caso? Basándonos en las pruebas que hemos revisado, una correlación de .40 es una estimación muy generosa. En consecuencia, una estimación regresiva de las ventas del primer año del nuevo contratado sería, como máximo, de 500.000 dólares + (1.000.000 - 500.000 dólares) × .40 = 700.000 dólares.

Este proceso no es, de nuevo, completamente intuitivo. En particular, como ilustran los ejemplos, las predicciones corregidas siempre serán más conservadoras que las intuitivas: nunca serán tan extremas como estas últimas, sino que se acercarán (a menudo «mucho más») a la media. Si corregimos nuestras predicciones, nunca apostaremos a que el campeón de tenis que ha ganado diez títulos de Grand Slam ganará otros diez. Tampoco pronosticaremos que una empresa de gran éxito valorada en mil millones de dólares se convertirá en un mastodonte que valga varios cientos de veces esa cifra. Las predicciones corregidas no aceptan apuestas sobre casos atípicos.

Esto significa que, retrospectivamente consideradas, las predicciones corregidas darán lugar de forma inevitable a algunos fallos muy visibles. Sin embargo, la predicción no se hace en retrospectiva. Debemos recordar que los casos atípicos son, por definición, rarísimos. El error contrario es

mucho más frecuente: cuando predecimos que los casos atípicos seguirán siendo atípicos, generalmente no lo serán debido a la regresión a la media. Por eso, cuando se maximiza la exactitud (es decir, se minimiza el ECM), las predicciones corregidas son superiores a las predicciones intuitivas emparejadoras.

PARA ENTENDER CÓMO PENSAMOS Y PENSAR MEJOR, HAY QUE LEER A KAHNEMAN

Premio Nobel de Economía

Dos médicos en la misma ciudad pueden dar diagnósticos diferentes a pacientes idénticos; dos jueces pueden dictar sentencias distintas ante delitos similares; nosotros mismos podemos decidir una cosa u otra según sea por la mañana o por la tarde, o según se acerque o no la hora de comer. Estos son ejemplos de ruido: el sesgo que conlleva variabilidad en juicios que deberían ser iguales.

El ruido está presente en todas las decisiones individuales y colectivas, y produce errores en innumerables terrenos, desde la medicina hasta la economía, pasando por el derecho, la sanidad, la protección infantil y la contratación. Además, también nos importuna e influye a la hora de tomar muchas de nuestras decisiones cotidianas.

Daniel Kahneman, uno de los psicólogos más importantes del mundo, junto con Cass R. Sunstein y Olivier Sibony, dos eminencias mundiales en pensamiento estratégico, nos enseñan a escuchar ese ruido, cuyo impacto tendemos a ignorar, y a reducirlo para mejorar nuestros juicios. Basado en

el mismo tipo de análisis agudo e ingeniosos ejemplos que convirtió Pensar rápido, pensar despacio en un best seller internacional, Ruido ofrece una serie de remedios originales, prácticos y sencillos para pensar mejor.

La crítica ha dicho:

«Una investigación absolutamente brillante de un problema social inmenso que nos ha pasado desapercibido cuando ha estado siempre ante nuestros ojos.»
Steven Levitt, autor de *Freakonomics*

«Monumental y apasionante. Excepcional.»
The Sunday Times

«Un tour de force de sabiduría y claridad.»
The New York Times

Daniel Kahneman es profesor emérito de Psicología y actividades públicas en la Universidad de Princeton y ganador del Premio Nobel de Ciencias Económicas del año 2002, así como de la Presidential Medal of Freedom en 2013. Kahneman es miembro de la American Academy of Arts and Sciences y de la National Academy of Sciences. También es socio de la American Psychological Association, la American Psychological Society, la Society of Experimental Psychologists y la Econometric Society. Ha recibido numerosos galardones, entre ellos el Distinguished Scientific Contribution Award de la American Psychological Association, la Warren Medal de la Society of Experimental Psychologists, el Hilgard Award for Career Contributions to General Psychology y el Award for Lifetime Contributions to Psychology que concede la American Psychological Association. Es autor del libro *Thinking, Fast and Slow* (*Pensar rápido, pensar despacio*, Barcelona, Debate, 2012), best seller de *The New York Times*. Vive en la ciudad de Nueva York.

Olivier Sibony es profesor de Estrategia en la Escuela de Estudios Superiores de Comercio (HEC) de París y miembro asociado de la Escuela de Negocios Saïd de la Universidad de Oxford. Anteriormente, pasó veinticinco años en las oficinas de McKinsey & Company en París y Nueva York, donde fue socio principal. Las investigaciones de Sibony sobre la mejora de la calidad de las decisiones estratégicas han aparecido en muchas publicaciones, como *Harvard Business Review* y *MIT Sloan Management Review*. Se graduó en la HEC de París y se doctoró en la Université Paris Sciences et Lettres. Es autor del libro *Vous allez commettre une terrible erreur!* (2019). Vive en París. Twitter: @siboliv.

Cass R. Sunstein es profesor en la Universidad Robert Walmsley de Harvard, donde es fundador y director del Program on Behavioral Economics and Public Policy. De 2009 a 2012 fue administrador de la Office of Information and Regulatory Affairs de la Casa Blanca. De 2013 a 2014 formó parte del Review Group on Intelligence and Communications Technologies del presidente Obama. Ganador en 2018 del Holberg Prize del

Gobierno de Noruega, Sunstein es autor de numerosos artículos y libros, entre ellos dos best sellers de *The New York Times*: *The World According to Star Wars* y *Nudge* (*Un pequeño empujón*, Barcelona, Taurus, 2017) en colaboración con Richard H. Thaler. Otros libros suyos son *How Change Happens* y *Too Much Information*. Twitter: @casssunstein.

Título original: *Noise. A Flaw in Human Judgment*

Edición en formato digital: septiembre de 2021

© 2021, Daniel Kahneman, Olivier Sibony y Cass R. Sunstein
Todos los derechos reservados
© 2021, Penguin Random House Grupo Editorial, S. A. U.
Travessera de Gràcia, 47-49. 08021 Barcelona
© 2021, Joaquín Chamorro Mielke, por la traducción

Diseño de portada: Adaptación del diseño original de Julia Humphreys. © HarperCollinsPublishers 2021

Penguin Random House Grupo Editorial apoya la protección del *copyright*. El *copyright* estimula la creatividad, defiende la diversidad en el ámbito de las ideas y el conocimiento, promueve la libre expresión y favorece una cultura viva. Gracias por comprar una edición autorizada de este libro y por respetar las leyes del *copyright* al no reproducir ni distribuir ninguna parte de esta obra por ningún medio sin permiso. Al hacerlo está respaldando a los autores y permitiendo que PRHGE continúe publicando libros para todos los lectores. Diríjase a CEDRO (Centro Español de Derechos Reprográficos, www.cedro.org) si necesita reproducir algún fragmento de esta obra.

ISBN: 978-84-1800-637-1

Composición digital: Newcomlab S.L.L.

Facebook: PenguinEbooks
Facebook: debatelibros
Twitter: @debatelibros
Instagram: @debatelibros
YouTube: penguinlibros
Spotify: PenguinLibros

Introducción. Dos clases de error

[1] Utilizando arcos y flechas en lugar de pistolas, el matemático suizo Daniel Bernoulli ofreció la misma analogía en 1778 en un ensayo sobre problemas de estimación. D. Bernoulli, «The Most Probable Choice Between Several Discrepant Observations and the Formation Therefrom of the Most Likely Induction», *Biometrika*, 48, 1-2 (junio de 1961), pp. 3-18, <https://doi.org/10.1093/biomet/48.1-2.3>.

[2] Joseph J. Doyle Jr., «Child Protection and Child Outcomes. Measuring the Effects of Foster Care», *American Economic Review*, 95, 5 (diciembre de 2007), pp. 1583-1610.

[3] Stein Grimstad y Magne Jørgensen, «Inconsistency of Expert Judgment-Based Estimates of Software Development Effort», *Journal of Systems and Software*, 80, 11 (2007), pp. 1770-1777.

[4] Andrew I. Schoenholtz, Jaya Ramji-Nogales y Philip G. Schrag, «Refugee Roulette. Disparities in Asylum Adjudication», *Stanford Law Review*, 60, 2 (2007).

[5] Mark A. Lemley y Bhaven Sampat, «Examiner Characteristics and Patent Office Outcomes», *Review of Economics and Statistics*, 94, 3 (2012), pp. 817-827. Véase también Iain Cockburn, Samuel Kortum y Scott Stern, «Are All Patent Examiners Equal? The Impact of Examiner Characteristics», documento de trabajo 8980, junio de 2002, <www.nber.org/papers/ w8980>, y Michael D. Frakes y Melissa F. Wasserman, «Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data», *Review of Economics and Statistics*, 99, 3 (julio de 2017), pp. 550-563.

1. Crimen y castigo ruidoso

[1] Marvin Frankel, *Criminal Sentences. Law Without Order*, 25 Inst. for Sci. Info. Current Contents / Soc. & Behavioral Scis.: This Week's Citation Classic 14, 2A-6 (23 de junio de 1986), <http://www.garfield.library. upenn.edu/classics1986/A1986C697400001.pdf>.

[2] Marvin Frankel, *Criminal Sentences. Law Without Order*, Nueva York, Hill & Wang, 1973, p. 5.

[3]. *Ibid*., p. 103.

[4]. *Ibid*., p. 5.

[5]. *Ibid*., p. 11.

[6]. *Ibid*., p. 114.

[7]. *Ibid*., p. 115.

[8]. *Ibid*., p. 119.

[9] Anthony Partridge y William B. Eldridge, *The Second Circuit Sentence Study. A Report to the Judges of the Second Circuit August 1974*, Washington DC, Federal Judicial Center, agosto de 1974, p. 9.

[10] Senado de Estados Unidos, «Comprehensive Crime Control Act of 1983. Report of the Committee on the Judiciary, United States Senate, on S. 1762, Together with Additional and Minority Views», Washington DC, US Government Printing Office, 1983, informe 98-225.

[11] A. Partridge y W. B. Eldridge, *The Second Circuit Sentence Study*..., A-11.

[12] *Ibid*., A-9.

[13] *Ibid*., A-5 - A-7.

[14] William Austin y Thomas A. Williams III, «A Survey of Judges. Responses to Simulated Legal Cases. Research Note on Sentencing Disparity», *Journal of Criminal Law & Criminology*, 68 (1977), p. 306.

[15] John Bartolomeo *et al*., «Sentence Decision Making. The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity», *Journal of Criminal Law and Criminology*, 72, 2 (1981). (Véase el capítulo 6 para una discusión más amplia.) Véase también Senado de Estados Unidos, p. 44.

[16] Shai Danziger, Jonathan Levav y Liora Avanim-Pesso, «Extraneous Factors in Judicial Decisions», *Proceedings of the National Academy of Sciences of the United States of America*, 108, 17 (2011), pp. 6889-6892.

[17] Ozkan Eren y Naci Mocan, «Emotional Judges and Unlucky Juveniles», *American Economic Journal. Applied Economics*, 10, 3 (2018), pp. 171205.

[18] Daniel L. Chen y Markus Loecher, «Mood and the Malleability of Moral Reasoning. The Impact of Irrelevant Factors on Judicial Decisions», *SSRN Electronic Journal* (21 de septiembre de 2019), p. 170, <http://users. nber.org/dlchen/papers/Mood_and_the_Malleability_of_Moral_Reasoning.pdf>.

[19] Daniel L. Chen y Arnaud Philippe, «Clash of Norms. Judicial Leniency on Defendant Birthdays», (2020), accesible en *SSRN*, <https://ssrn. com/abstract=3203624>.

[20] Anthony Heyes y Sooedeh Saberian, «Temperature and Decisions. Evidence from 207,000 Court Cases», *American Economic Journal. Applied Economics*, 11, 2 (2018), pp. 238-265.

[21] Senado de Estados Unidos, p. 38.

[22] Senado de Estados Unidos, p. 38.

[23]. Las palabras del juez Breyer aparecen citadas en Jeffrey Rosen, «Breyer Restraint», *The New Republic*, 11 de julio de 1994, pp. 19 y 25.

[24] Comisión de Sentencias de Estados Unidos, «Manual de directrices» (2018), en <www.ussc.gov/sites/default/files/pdf/guidelines-manual/ 2018/GLMFull.pdf>.

[25] James M. Anderson, Jeffrey R. Kling y Kate Stith, «Measuring Interjudge Sentencing Disparity. Before and After the Federal Sentencing Guidelines», *Journal of Law and Economics*, 42, S1 (abril de 1999), pp. 271-308.

[26] Comisión de Sentencias de Estados Unidos, *The Federal Sentencing Guidelines. A Report on the Operation of the Guidelines System and Short-Term Impacts on Disparity in Sentencing, Use of Incarceration, and Prosecutorial Discretion and Plea Bargaining*, vols. 1 y 2, Washington DC, 1991.

[27] J. M. Anderson, J. R. Kling y K. Stith, «Measuring Interjudge Sentencing Disparity...».

[28] Paul J. Hofer, Kevin R. Blackwell y R. Barry Ruback, «The Effect of the Federal Sentencing Guidelines on Inter-Judge Sentencing Disparity», *Journal of Criminal Law and Criminology*, 90 (1999), pp. 239 y 241.

[29] Kate Stith y José Cabranes, *Fear of Judging. Sentencing Guidelines in the Federal Courts*, Chicago, University of Chicago Press, 1998, p. 79.

[30] 543 U. S. 220 (2005).

[31] Comisión de Sentencias de Estados Unidos, «Results of Survey of United States District Judges, January 2010 through March 2010», cuestión 19, tabla 19, junio de 2010, <www.ussc.gov/sites/default/files/pdf/re search-and-publications/research-projects-and-surveys/surveys/2010 0608_Judge_Survey.pdf>.

[32] Crystal Yang, «Have Interjudge Sentencing Disparities Increased in an Advisory Guidelines Regime? Evidence from Booker», *New York University Law Review*, 89 (2014), pp. 1268-1342; pp.

1278 y 1334.

2. Un sistema ruidoso

[1] Los ejecutivos de la compañía hicieron descripciones detalladas de casos representativos, similares a los riesgos y reclamaciones con que los empleados trataban a diario. Se prepararon seis casos para los tasadores de reclamaciones de la división de bienes y accidentes, y cuatro para los agentes especializados en riesgos financieros. A los empleados se les dio medio día libre de su trabajo habitual para evaluar dos o tres casos cada uno. Se les instruyó para que trabajaran de forma independiente y no se les dijo que el propósito del estudio era examinar la variabilidad de sus juicios. Obtuvimos un total de seis-ocho juicios de cuarenta y ocho agentes y ciento trece juicios de sesenta y ocho tasadores de reclamaciones.

[2] Dale W. Griffin y Lee Ross, «Subjective Construal, Social Inference, and Human Misunderstanding», *Advances in Experimental Social Psychology*, 24 (1991), pp. 319-359; Robert J. Robinson, Dacher Keltner, Andrew Ward y Lee Ross, «Actual Versus Assumed Differences in Construal: "Naive Realism" in Intergroup Perception and Conflict», *Journal of Personality and Social Psychology*, 68, 3 (1995), p. 404; Lee Ross y Andrew Ward, «Naive Realism in Everyday Life. Implications for Social Conflict and Misunderstanding», en *Values and Knowledge*, Edward S. Reed, Elliot Turiel y Terrance Brown (eds.), Hillsdale (New Jersey), Lawrence Erlbaum Associates, 1996, pp. 103-135.

[1] La desviación estándar de un conjunto de números se deriva de otra cantidad estadística, llamada «varianza». Para calcular esta, primero se obtiene la distribución de las desviaciones respecto de la media, y luego se toma el cuadrado de cada una de estas desviaciones. La varianza es la media de estas desviaciones elevadas al cuadrado, y la desviación estándar es la raíz cuadrada de la varianza.

4. Cuestiones de juicio

[1] R. T. Hodgson, «An Examination of Judge Reliability at a Major U. S. Wine Competition», *Journal of Wine Economics*, 3, 2 (2008), pp. 105-113.

[2] Algunos estudiosos de la toma de decisiones definen las «decisiones» como elecciones entre opciones, y consideran que los juicios cuantitativos son un caso especial de decisión en el que hay un continuo de opciones posibles. Desde este punto de vista, los juicios son un caso especial de decisión. Nuestro planteamiento es aquí diferente: consideramos que las decisiones que exigen una elección entre opciones se derivan de un juicio evaluativo subyacente sobre cada opción. Es decir, consideramos las decisiones como un caso especial de juicio.

5. La medición del error

[1] El método de los mínimos cuadrados fue publicado por primera vez por Adrien-Marie Legendre en 1805. Gauss afirmó que lo había utilizado por primera vez diez años antes, y

posteriormente lo relacionó con el desarrollo de una teoría del error y con la curva de error normal, que lleva su nombre. El conflicto de prelación ha sido muy debatido, y los historiadores se inclinan por creer la afirmación de Gauss (Stephen M. Stigler, «Gauss and the Invention of Least Squares», *Annals of Statistics*, 9 [1981], pp. 465-474; y Stephen M. Stigler, *The History of Statistics. The Measurement of Uncertainty Before 1900*, Cambridge, Massachusetts, Belknap Press of Harvard University Press, 1986).

[2] Hemos definido el «ruido» como la desviación estándar de los errores; por tanto, el ruido al cuadrado es la varianza de los errores. La definición de «varianza» es «la media de los cuadrados menos el cuadrado de la media». Como el error medio es el sesgo, el «cuadrado de la media» es el sesgo al cuadrado. Por tanto: ruido^2 = ECM – sesgo^2.

[3] Berkeley J. Dietvorst y Soaham Bharti, «People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error», *Psychological Science*, 31, 10 (2020), pp. 1302-1314.

6. El análisis del ruido

[1] John Bartolomeo, Kevin Clancy, David Richardson y Charles Wellford, «Sentence Decision Making. The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity, *Journal of Criminal Law and Criminology*, 72, 2 (1981), pp. 524-554; INSLAW, Inc. *et al*., «Federal Sentencing. Towards a More Explicit Policy of Criminal Sanctions III-4», 1981.

[2] La sentencia podía incluir cualquier combinación de tiempo en prisión, tiempo supervisado y multas. Para simplificar, nos centramos aquí sobre todo en el componente principal de las sentencias —el tiempo de prisión— y dejamos de lado los otros dos.

[3] En un entorno de casos múltiples y jueces múltiples, la versión ampliada de la ecuación de error que introdujimos en el capítulo 5 incluye un término que refleja esta varianza. En concreto, si definimos un «sesgo general» como el error medio de todos los casos, y si este error no es idéntico en todos los casos, habrá una varianza de los sesgos de los casos. La ecuación se convierte en:ECM = sesgo general^2 + varianza de sesgos de casos + ruido del sistema^2.

[4] Las cifras mencionadas en este capítulo se derivan del estudio original de la siguiente manera:

En primer lugar, los autores señalan que el efecto principal «del delito y del delincuente» representa el 45 por ciento de la varianza total (John Bartolomeo *et al*., «Sentence Decision Making...», tabla 6). Sin embargo, aquí nos ocupamos más ampliamente del efecto de cada caso, incluyendo todas las características presentadas a los jueces —como si el acusado tenía antecedentes penales o si utilizó un arma en la comisión del delito—. Según nuestra definición, todas estas características forman parte de la «verdadera varianza de los casos», no del ruido. En consecuencia, reintegramos las interacciones entre las características de cada caso en la varianza del caso (estas representan el 11 por ciento de la varianza total; véase *Ibid*., tabla 10). Como resultado, redefinimos la parte de la varianza del caso como del 56 por ciento, la del efecto principal del juez (ruido de nivel) como del 21

por ciento y la de las interacciones en la varianza total como del 23 por ciento. El ruido del sistema es, por tanto, el 44 por ciento de la varianza total.

La varianza de las sentencias justas puede calcularse a partir de *Ibid*., p. 89, en la tabla que recoge las sentencias medias para cada caso: la varianza es de 15. Si esto es el 56 por ciento de la varianza total, la varianza total es de 26,79, y la varianza de ruido del sistema es de 11,79. La raíz cuadrada de esa varianza es la desviación típica de un caso representativo, que son 3,4 años.

El efecto principal del juez, o ruido de nivel, es el 21 por ciento de la varianza total. La raíz cuadrada de esa varianza es la desviación estándar atribuible al ruido de nivel del juez, que son 2,4 años.

[5] Este valor es la raíz cuadrada de la media de las varianzas de las sentencias aplicadas a los dieciséis casos. Lo hemos calculado como se explica en la nota anterior.

[6] La hipótesis de la aditividad supone, efectivamente, que la dureza de un juez añade una cantidad constante de tiempo de prisión. Es poco probable que esta hipótesis sea correcta: es más probable que la dureza del juez añada una cantidad proporcional a la sentencia media. Esta cuestión fue ignorada en el informe original, que no permite evaluar su importancia.

[7] J. Bartolomeo *et al*., «Sentence Decision Making...», p. 23.

[8] Se cumple la siguiente ecuación: $(\text{ruido del sistema})^2 = (\text{ruido de nivel})^2 + (\text{ruido de patrón})^2$. La tabla muestra que el ruido del sistema es de 3,4 años, y que el ruido de nivel es de 2,4 años. De ello se sigue que el ruido de patrón es también de unos 2,4 años. El cálculo se muestra a modo de ilustración; los valores reales son ligeramente diferentes debido a errores de redondeo.

7. Ruido de ocasión

[1]. Véase <http://www.iweblists.com/sports/basketball/FreeThrow Percent_c.html>, consultada el 27 de diciembre de 2020.

[2] Véase <https://www.basketball-reference.com/players/o/onealsh 01.html>, consultada el 27 de diciembre de 2020.

[3] R. T. Hodgson, «An Examination of Judge Reliability at a Major U.S. Wine Competition», *Journal of Wine Economics*, 3, 2 (2008), pp. 105-113.

[4] Stein Grimstad y Magne Jørgensen, «Inconsistency of Expert Judgment-Based Estimates of Software Development Effort», *Journal of Systems and Software*, 80, 11 (2007), pp. 1770-1777.

[5] Robert H. Ashton, «A Review and Analysis of Research on the Test-Retest Reliability of Professional Judgment», *Journal of Behavioral Decision Making*, 294, 3 (2000), pp. 277-294. Por cierto, el autor señaló entonces que ni uno solo de los cuarenta y un estudios que revisó fue diseñado para evaluar el ruido de ocasión: «En todos los casos, la medición de la fiabilidad fue un subproducto de algunos otros objetivos de la investigación», *Ibid*., p. 279. Este comentario indica que el interés por estudiar el ruido de ocasión es relativamente nuevo.

[6] Central Intelligence Agency (CIA), *The World Factbook*, Washington DC, 2020. La cifra citada incluye todos los aeropuertos o aeródromos reconocibles desde el aire. La pista, o las pistas, pueden estar pavimentadas o no, y pueden incluir instalaciones cerradas o abandonadas.

[7] Edward Vul y Harold Pashler, «Measuring the Crowd Within. Probabilistic Representations Within Individuals», *Psychological Science*, 19, 7 (2008), pp. 645-648.

[8] James Surowiecki, *The Wisdom of Crowds. Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*, Nueva York, Doubleday, 2004.

[9] La desviación estándar de los juicios promediados (nuestra medida del ruido) decrece en proporción a la raíz cuadrada del número de juicios.

[10] E. Vul y H. Pashler, «Measuring the Crowd Within...», p. 646.

[11]. Stefan M. Herzog y Ralph Hertwig, «Think Twice and Then. Combining or Choosing in Dialectical Bootstrapping?», *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 40, 1 (2014), pp. 218-232.

[12] E. Vul y H. Pashler, «Measuring the Crowd Within...», p. 647.

[13] Joseph P. Forgas, «Affective Influences on Interpersonal Behavior», *Psychological Inquiry*, 13, 1 (2002), pp. 1-28.

[14] *Ibid*., p. 10.

[15] A. Filipowicz, S. Barsade y S. Melwani, «Understanding Emotional Transitions: the Interpersonal Consequences of Changing Emotions in Negotiations», *Journal of Personality and Social Psychology*, 101, 3 (2011), pp. 541-556.

[16] Joseph P. Forgas, «She Just Doesn't Look like a Philosopher...? Affective Influences on the Halo Effect in Impression Formation», *European Journal of Social Psychology*, 41, 7 (2011), pp. 812-817.

[17] Gordon Pennycook, James Allan Cheyne, Nathaniel Barr, Derek J. Koehler y Jonathan A. Fugelsang, «On the Reception and Detection of PseudoProfound Bullshit», *Judgment and Decision Making*, 10, 6 (2015), pp. 549-563.

[18] Harry Frankfurt, *On Bullshit*, Princeton, New Jersey, Princeton University Press, 2005.

[19] G. Pennycook *et al*., «On the Reception and Detection...», p. 549.

[20] Joseph P. Forgas, «Happy Believers and Sad Skeptics? Affective Influences on Gullibility», *Current Directions in Psychological Science*, 28, 3 (2019), pp. 306-313.

[21] Joseph P. Forgas, «Mood Effects on Eyewitness Memory. Affective Influences on Susceptibility to Misinformation», *Journal of Experimental Social Psychology*, 41, 6 (2005), pp. 574-588.

[22] Piercarlo Valdesolo y David Desteno, «Manipulations of Emotional Context Shape Moral Judgment», *Psychological Science*, 17, 6 (2006), pp. 476-477.

[23] Hannah T. Neprash y Michael L. Barnett, «Association of Primary Care Clinic Appointment Time with Opioid Prescribing», *JAMA Network Open*, 2, 8 (2019); Lindsey M. Philpot, Bushra A. Khokhar, Daniel L. Roellinger, Priya Ramar y Jon O. Ebbert, «Time of Day Is Associated with Opioid Prescribing for Low Back Pain in Primary Care», *Journal of General Internal Medicine*, 33 (2018), p. 1828.

[24] Jeffrey A. Linder, Jason N. Doctor, Mark W. Friedberg, Harry Reyes Nieva, Caroline Birks, Daniella Meeker y Craig R. Fox, «Time of Day and the Decision to Prescribe Antibiotics», *JAMA Internal Medicine*, 174, 12 (2014), pp. 2029-2031.

[25] Rebecca H. Kim, Susan C. Day, Dylan S. Small, Christopher K. Snider, Charles A. L. Rareshide y Mitesh S. Patel, «Variations in Influenza Vaccination by Clinic Appointment Time and

an Active Choice Intervention in the Electronic Health Record to Increase Influenza Vaccination», *JAMA Network Open*, 1, 5 (2018), pp. 1-10.

[26] Sobre la mejora de la memoria, véase Joseph P. Forgas, Liz Goldenberg y Christian Unkelbach, «Can Bad Weather Improve Your Memory? An Unobtrusive Field Study of Natural Mood Effects on Real-Life Memory», *Journal of Experimental Social Psychology*, 45, 1 (2008), pp. 254-257. Sobre el efecto de la luz solar, véase David Hirshleifer y Tyler Shumway, «Good Day Sunshine. Stock Returns and the Weather», *Journal of Finance*, 58, 3 (2003), pp. 1009-1032.

[27] Uri Simonsohn, «Clouds Make Nerds Look Good. Field Evidence of the Impact of Incidental Factors on Decision Making», *Journal of Behavioral Decision Making*, 20, 2 (2007), pp. 143-152.

[28] Daniel Chen *et al*., «Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires», *Quarterly Journal of Economics*, 131, 3 (2016), pp. 1181-1242.

[29] Jaya Ramji-Nogales, Andrew I.Schoenholtz y Philip Schrag,«Refugee Roulette. Disparities in Asylum Adjudication», *Stanford Law Review*, 60, 2 (2007).

[30] Michael J. Kahana *et al*., «The Variability Puzzle in Human Memory», *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 44, 12 (2018), pp. 1857-1863.

8. Cómo los grupos amplifican el ruido

[1] Matthew J. Salganik, Peter Sheridan Dodds y Duncan J. Watts, «Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market», *Science*, 311 (2006), pp. 854-856. Véanse también Matthew Salganik y Duncan Watts, «Leading the Herd Astray. An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market», *Social Psychology Quarterly*, 71 (2008), pp. 338-355, y Matthew J. Salganik y Duncan J. Watts, «Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets», *Topics in Cognitive Science*, 1 (2009), pp. 439-468.

[2] M. Salganik y D. Watts, «Leading the Herd Astray...», p. 97.

[3] Michael Macy *et al*., «Opinion Cascades and the Unpredictability of Partisan Polarization», *Science Advances*, (2019), pp. 1-8. Véase también Helen Margetts *et al*., *Political Turbulence*, Princeton, Princeton University Press (2015).

[4] M. Macy *et al*., «Opinion Cascades...».

[5] Lev Muchnik *et al*., «Social Influence Bias. A Randomized Experiment», *Science*, 341, 6146 (2013), pp. 647-651.

[6] Jan Lorenz *et al*., «How Social Influence Can Undermine the Wisdom of Crowd Effect», *Proceedings of the National Academy of Sciences*, 108, 22 (2011), pp. 9020-9025.

[7] Daniel Kahneman, David Schkade y Cass Sunstein, «Shared Outrage and Erratic Awards. The Psychology of Punitive Damages», *Journal of Risk and Uncertainty*, 16 (1998), pp. 49-86.

[8] David Schkade, Cass R. Sunstein y Daniel Kahneman, «Deliberating about Dollars. The Severity Shift», *Columbia Law Review*, 100 (2000), pp. 1139-1175.

[1] El porcentaje de concordancia (PC) está estrechamente relacionado con la prueba W de Kendall, también conocida como «coeficiente de concordancia».

[2] Kanwal Kamboj *et al*., «A Study on the Correlation Between Foot Length and Height of an Individual and to Derive Regression Formulae to Estimate the Height from Foot Length of an

Individual», *International Journal of Research in Medical Sciences*, 6, 2 (2018), p. 528.

[3] El PC se calcula suponiendo que la distribución conjunta es bivariante-normal. Los valores indicados en la tabla son aproximaciones basadas en esta suposición. Agradecemos a Julian Parris la elaboración de esta tabla.

9. Juicios y modelos

[1] Martin C. Yu y Nathan R. Kuncel, «Pushing the Limits for Judgmental Consistency. Comparing Random Weighting Schemes with Expert Judgments», *Personnel Assessment and Decisions*, 6, 2 (2020), pp. 1-10. La correlación de .15 obtenida por los expertos es la media no ponderada de las tres muestras estudiadas, que incluyen ochocientos cuarenta y siete casos en total. El estudio real difiere de esta descripción simplificada en varios aspectos.

[2] Un requisito previo para construir una media ponderada es que todos los predictores deben medirse en unidades comparables. Este requisito se cumplió en nuestro ejemplo de la introducción, en el que todas las valoraciones se hicieron en una escala de 0 a 10, pero no siempre es así. Por ejemplo, los predictores del rendimiento podrían ser la evaluación del entrevistador en una escala de 0 a 10, el número de años de experiencia pertinente y la puntuación en un test de aptitud. El programa de regresión múltiple transforma a todos los predictores en «puntuaciones estándar» antes de combinarlos. Una puntuación estándar mide la distancia de una observación respecto de la media de una población, con la desviación estándar como unidad. Por ejemplo, si la media de la prueba de aptitud es 55 y la desviación estándar es 8, una puntuación estándar de +1,5 corresponde a un resultado de 67 en el test. En particular, la estandarización de los datos de cada individuo elimina cualquier rastro de error en la media o en la varianza de los juicios de los individuos.

[3] Una característica importante de la regresión múltiple es que el peso óptimo de cada predictor depende de los demás predictores. Si un predictor está muy correlacionado con otro, no debería tener una ponderación igual de grande (esto sería una forma de «doble contabilidad»).

[4] Robin M. Hogarth y Natalia Karelaia, «Heuristic and Linear Models of Judgment. Matching Rules and Environments», *Psychological Review*, 114, 3 (2007), p. 734.

[5] Un marco de investigación que se ha utilizado ampliamente en este contexto es el «modelo de lente de los juicios», en el que se basa esta discusión. Véanse Kenneth R. Hammond, «Probabilistic Functioning and the Clinical Method», *Psychological Review*, 62, 4 (1955), pp. 255-262, y Natalia Karelaia y Robin M. Hogarth, «Determinants of Linear Judgment. A Meta-Analysis of Lens Model Studies», *Psychological Bulletin*, 134, 3 (2008), pp. 404-426.

[6] Paul E. Meehl, *Clinical Versus Statistical Prediction. A Theoretical Analysis and a Review of the Evidence*, Minneapolis, University of Minnesota Press, 1954.

[7] Paul E. Meehl, *Clinical Versus Statistical Prediction. A Theoretical Analysis and a Review of the Evidence*, Northvale, New Jersey, Aronson, 1996, prólogo.

[8] «Paul E.Meehl», en Ed Lindzey (ed.), *A History of Psychology in Autobiography*, Washington D. C., American Psychological Association, 1989.

[9]. *Ibid*., p. 362.

[10] William M. Grove *et al*., «Clinical Versus Mechanical Prediction. A Meta-Analysis», *Psychological Assessment*, 12, 1 (2000), pp. 19-30.

[11] William M. Grove y Paul E. Meehl, «Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures. The Clinical-Statistical Controversy», *Psychology, Public Policy, and Law*, 2, 2 (1996), pp. 293-323.

[12] Lewis Goldberg, «Man Versus Model of Man. A Rationale, plus Some Evidence, for a Method of Improving on Clinical Inferences», *Psychological Bulletin*, 73, 6 (1970), pp. 422-432.

[13] Milton Friedman y Leonard J. Savage, «The Utility Analysis of Choices Involving Risk», *Journal of Political Economy*, 56, 4 (1948), pp. 279304.

[14] N. Karelaia y R. M. Hogarth, «Determinants of Linear Judgment...», p. 411, tabla 1.

[15] Nancy Wiggins y Eileen S. Kohen, «Man Versus Model of Man Revisited. The Forecasting of Graduate School Success», *Journal of Personality and Social Psychology*, 19, 1 (1971), pp. 100-106.

[16] N. Karelaia y R. M. Hogarth, «Determinants of Linear Judgment...».

[17] La corrección de un coeficiente de correlación por la fiabilidad imperfecta del predictor se conoce como «corrección por atenuación». La fórmula es r_{xy} corregido = $r_{xy}/\sqrt{}$ r_{xx}, donde r_{xx} es el coeficiente de fiabilidad (la proporción de la varianza verdadera en la varianza observada del predictor).

[18] M. C. Yu y N. R. Kuncel, «Pushing the Limits for Judgmental Consistency...».

[19] En el próximo capítulo analizaremos con más detalle los modelos de peso igual y de peso aleatorio. Los pesos están limitados a un rango de números pequeños, y han de tener el signo correcto.

10. Reglas sin ruido

[1] Robyn M. Dawes y Bernard Corrigan, «Linear Models in Decision Making», *Psychological Bulletin*, 81, 2 (1974), pp. 95-106. Dawes y Corrigan también propusieron utilizar pesos aleatorios. El estudio de la predicción del rendimiento de directivos descrito en el capítulo 9 es una aplicación de esta idea.

[2] Jason Dana, «What Makes Improper Linear Models Tick?», en *Rationality and Social Responsibility. Essays in Honor of Robyn M. Dawes*, Joachim I. Krueger (ed.), pp. 71-89, Nueva York, Psychology Press, 2008, p. 73.

[3] Jason Dana y Robyn M. Dawes, «The Superiority of Simple Alternatives to Regression for Social Sciences Prediction», *Journal of Educational and Behavior Statistics*, 29 (2004), pp. 317-331; J. Dana, «What Makes Improper Linear Models Tick?», en *Rationality and Social Responsibility*...

[4] Howard Wainer, «Estimating Coefficients in Linear Models. It Don’t Make No Nevermind», Psychological Bulletin, 83, 2 (1976), pp. 213-217.

[5] J. Dana, «What Makes Improper Linear Models Tick?», en *Rationality and Social Responsibility*..., p. 72.

[6] Martin C. Yu y Nathan R. Kuncel, «Pushing the Limits for Judgmental Consistency. Comparing Random Weighting Schemes with Expert Judgments», *Personnel Assessment and Decisions*, 6,2 (2020), pp. 1-10. Al igual que en el capítulo anterior, la correlación señalada es la media no ponderada de las tres muestras estudiadas. La comparación se mantiene en cada una de las tres muestras: la validez del juicio clínico de los expertos fue de .17, .16 y .13, y la validez de los modelos de peso igual fue de .19, .33 y .22, respectivamente.

[7] Robyn M. Dawes, «The Robust Beauty of Improper Linear Models in Decision Making», *American Psychologist*, 34, 7 (1979), pp. 571-582.

[8] R. M. Dawes y B. Corrigan, «Linear Models in Decision Making»..., p. 105.

[9] Jongbin Jung, Conner Concannon, Ravi Shroff, Sharad Goel y Daniel G. Goldstein, «Simple Rules to Guide Expert Classifications», *Journal of the Royal Statistical Society, Statistics in Society*,

183 (2020), pp. 771-800.

[10] Julia Dressel y Hany Farid, «The Accuracy, Fairness, and Limits of Predicting Recidivism», *Science Advances*, 4, 1 (2018), pp. 1-6.

[11] Estos dos ejemplos son modelos lineales basados en un conjunto extremadamente pequeño de variables (y, en el caso del modelo de la libertad bajo fianza, en una aproximación de los pesos lineales obtenida por un método de redondeo que transforma el modelo en un cálculo de servilleta). Otro tipo de «modelo impropio» es una regla de «una sola variable» que considera solo un predictor e ignora todos los demás. Véase Peter M. Todd y Gerd Gigerenzer, «Précis of Simple Heuristics That Make Us Smart», *Behavioral and Brain Sciences*, 23, 5 (2000), pp. 727-741.

[12] P.Gendreau, T. Little y C.Goggin, «A Meta-Analysis of the Predictors of Adult Offender Recidivism. What Works!», *Criminology*, 34 (1996).

[13] En este contexto, el tamaño debe entenderse como la ratio entre el número de observaciones y los predictores. La «belleza sólida» de R. M. Dawes sugería que debe ser tan alta como de 15 o 20 a 1 para que las ponderaciones óptimas den un mejor resultado en la validación cruzada que las ponderaciones unitarias. En «Superiority of Simple Alternatives...», J. Dana y R. M. Dawes, utilizando muchos más estudios de casos, elevaron el listón a una ratio de 100 a 1.

[14] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig y S. Mullainathan, «Human Decisions and Machine Predictions», *Quarterly Journal of Economics*, 133 (2018), pp. 237-293.

[15] El algoritmo fue entrenado con un subconjunto de datos de entrenamiento, y luego se evaluó su capacidad para predecir resultados en un subconjunto diferente elegido al azar.

[16] J. Kleinberg *et al*., «Human Decisions...», p. 16.

[17] Gregory Stoddard, Jens Ludwig y Sendhil Mullainathan, correos electrónicos intercambiados con los autores de este libro en junio-julio de 2020.

[18] B. Cowgill, «Bias and Productivity in Humans and Algorithms. Theory and Evidence from Résumé Screening», artículo presentado en la Smith Entrepreneurship Research Conference, College Park, MD, 21 de abril de 2018.

[19] William M. Grove y Paul E. Meehl, «Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures. The Clinical-Statistical Controversy», *Psychology, Public Policy, and Law*, 2, 2 (1996), pp. 293-323.

[20] Jennifer M. Logg, Julia A. Minson y Don A. Moore, «Algorithm Appreciation. People Prefer Algorithmic to Human Judgment», *Organizational Behavior and Human Decision Processes*, 151 (abril de 2018), pp. 90-103.

[21] B. J. Dietvorst, J. P. Simmons y C. Massey, «Algorithm Aversion. People Erroneously Avoid Algorithms After Seeing Them Err», *Journal of Experimental Psychology General*, 144 (2015), pp. 114-126. Véase también A. Prahl y L. Van Swol, «Understanding Algorithm Aversion. When Is Advice from Automation Discounted?», *Journal of Forecasting*, 36 (2017), pp. 691-702.

[22] M. T. Dzindolet, L. G. Pierce, H. P. Beck y L. A. Dawe, «The Perceived Utility of Human and Automated Aids in a Visual Detection Task», *Human Factors. The Journal of the Human Factors and Ergonomics Society*, 44, 1 (2002), pp. 79-94; K. A. Hoffy M. Bashir, «Trust in Automation. Integrating Empirical Evidence on Factors That Influence Trust», *Human Factors. The Journal of the Human Factors and Ergonomics Society*, 57, 3 (2015), pp. 407434; P. Madhavan y D. A. Wiegmann, «Similarities and Differences Between Human-Human and Human-Automation Trust. An Integrative Review», *Theoretical Issues in Ergonomics Science*, 8, 4 (2007), pp. 277-301.

11. Ignorancia objetiva

[1] E. Dane y M. G. Pratt, «Exploring Intuition and Its Role in Managerial Decision Making», *Academy of Management Review*, 32, 1 (2007), pp. 33-54; Cinla Akinci y Eugene Sadler-Smith, «Intuition in Management Research. A Historical Review», *International Journal of Management Reviews*, 14 (2012), pp. 104-122; Gerard P. Hodgkinson *et al.*, «Intuition in Organizations. Implications for Strategic Management», *Long Range Planning*, 42 (2009), pp. 277-297.

[2]. *Ibid*., p. 279.

[3] Nathan Kuncel *et al*., «Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions. A Meta-Analysis», *Journal of Applied Psychology*, 98, 6 (2013), pp. 1060-1072. Para más consideraciones sobre las decisiones de selección de personal, véase también el capítulo 24.

[4] Don A. Moore, *Perfectly Confident. How to Calibrate Your Decisions Wisely*, Nueva York, HarperCollins, 2020.

[5] Philip E. Tetlock, *Expert Political Judgment. How Good Is It? How Can We Know*?, Princeton, New Jersey, Princeton University Press, 2005, pp. 239 y 233.

[6] William M. Grove *et al*., «Clinical Versus Mechanical Prediction. A Meta-Analysis», *Psychological Assessment*, 12, 1 (2000), pp. 19-30.

[7] Sendhil Mullainathan y Ziad Obermeyer, «Who Is Tested for Heart Attack and Who Should Be. Predicting Patient Risk and Physician Error», 2019, NBER documento de trabajo 26168, National Bureau of Economic Research.

[8] Weston Agor, «The Logic of Intuition. How Top Executives Make Important Decisions», *Organizational Dynamics*, 14, 3 (1986), pp. 5-18; Lisa A. Burke y Monica K. Miller, «Taking the Mystery Out of Intuitive Decision Making», *Academy of Management Perspectives*, 13, 4 (1999), pp. 91-99.

[9] Poornima Madhavan y Douglas A. Wiegmann, «Effects of Information Source, Pedigree, and Reliability on Operator Interaction with Decision Support Systems», *Human Factors. The Journal of the Human Factors and Ergonomics Society*, 49, 5 (2007).

12. El valle de lo normal

[1] Matthew J. Salganik *et al*., «Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration», *Proceedings of the National Academy of Sciences*, 117, 15 (2020), pp. 8398-8403.

[2] E incluían a 4.242 familias, ya que algunas de las familias del «Fragile Families and Child Wellbeing» fueron excluidas de este análisis por cuestiones de privacidad.

[3] Para puntuar la exactitud, los organizadores de la competición utilizaron la misma métrica que introdujimos en la primera parte: el error al cuadrado medio o ECM. Para facilitar la comparabilidad, también compararon el ECM de cada modelo con una estrategia de predicción «inútil»: una predicción igual para todos de que cada caso individual no es diferente de la media del conjunto de entrenamiento. Por comodidad, hemos convertido sus resultados en coeficientes de correlación. El ECM y la correlación están relacionados por la expresión r^2 = (Var [Y] – MSE) / Var (Y), donde Var (Y) es la varianza de la variable de resultado y (Var [Y] – MSE) es la varianza de los resultados predichos.

[4] F. D. Richard *et al*., «One Hundred Years of Social Psychology Quantitatively Described», *Review of General Psychology*, 7, 4 (2003), pp. 331663.

[5] Gilles E.Gignac y Eva T. Szodorai, «Effect Size Guidelines for Individual Differences Researchers», *Personality and Individual Differences*, 102 (2016), pp. 74-78.

[6] Es preciso hacer una advertencia. Por su planteamiento, este estudio utiliza un conjunto de datos descriptivos existente que es muy amplio, pero que no está específicamente adaptado para predecir resultados concretos. Esta es una diferencia importante con respecto a los expertos del estudio de Tetlock, que eran libres de utilizar cualquier información que consideraran oportuna. Es posible, por ejemplo, identificar predictores de desahucio que no están en la base de datos, pero que podrían reunirse. Por lo tanto, el estudio no demuestra lo «intrínsecamente» impredecibles que son los desahucios y otros resultados, sino lo impredecibles que son «sobre la base de este conjunto de datos», que es utilizado por numerosos científicos sociales.

[7] Jake M. Hofman *et al*., «Prediction and Explanation in Social Systems», *Science*, 355 (2017), pp. 486-488; Duncan J. Watts *et al*., «Explanation, Prediction, and Causality. Three Sides of the Same Coin?» (octubre de 2018), pp. 1-14, accesible en el Center for Open Science, <https://osf.io/bgwjc>.

[8]. Una distinción estrechamente relacionada pone en contraste el pensamiento «extensional» con el «no extensional» o «intencional». Amos Tversky y Daniel Kahneman, «Extensional Versus Intuitive Reasoning. The Conjunction Fallacy in Probability Judgment», *Psychological Review*, 4 (1983), pp. 293-315.

[9] Daniel Kahneman y Dale T. Miller, «Norm Theory. Comparing Reality to Its Alternatives», *Psychological Review*, 93, 2 (1986), pp. 136-153.

[10] Baruch Fischhoff, «An Early History of Hindsight Research», *Social Cognition*, 25, 1 (2007), pp. 10-13, doi:10.1521/soco.2007.25.1.10; Baruch Fischhoff, «Hindsight Is Not Equal to Foresight. The Effect of Outcome Knowledge on Judgment Under Uncertainty», *Journal of Experimental Psychology. Human Perception and Performance*, 1, 3 (1975), p. 288.

[11] Daniel Kahneman, *Thinking, Fast and Slow*, Nueva York, Farrar, Straus and Giroux, 2011. (Hay trad. cast.:*Pensar rápido, pensar despacio*, Barcelona, Debate, 2012.)

13. Heurísticas, sesgos y ruido

[1] Daniel Kahneman, *Thinking, Fast and Slow*, Nueva York, Farrar, Straus and Giroux, 2011. (Hay trad. cast.:*Pensar rápido, pensar despacio*, Barcelona, Debate, 2012.)

[2] Es preciso hacer una advertencia. Los psicólogos que estudian los sesgos de los juicios no se conforman con cinco participantes en cada grupo, como se muestra en la figura 10, por una muy buena razón: como los juicios son ruidosos, los resultados de cada grupo experimental rara vez se agruparán tan estrechamente como sugiere la figura 11. Las personas varían en su susceptibilidad a cada sesgo, y no ignoran «completamente» variables relevantes. Por ejemplo, con un número muy grande de participantes, se podría confirmar casi con certeza que la insensibilidad al alcance es imperfecta: la probabilidad media asignada a que Gambardi deje el puesto es muy ligeramente superior para tres años que para dos. Aun así, la descripción de la insensibilidad al alcance es apropiada porque la diferencia es una pequeña fracción de lo que debería ser.

[3] Daniel Kahneman *et al*. (eds.), *Judgment Under Uncertainty. Heuristics and Biases*, Nueva York, Cambridge University Press, 1982, cap. 6; Daniel Kahneman y Amos Tversky, «On the Psychology of Prediction», *Psychological Review*, 80, 4 (1973), pp. 237-251.

[4]. Véase, por ejemplo, Steven N. Kaplan y Bernadette A. Minton, «How Has CEO Turnover Changed?», *International Review of Finance*, 12, 1 (2012), pp. 57-87. Véase también Dirk Jenter y Katharina Lewellen, «Performance-Induced CEO Turnover», Harvard Law School Forum on Corporate Governance, 2 de septiembre de 2020, <https://corpgov.law.harvard.edu/2020/09/02/performance-induced-ceo-turnover>.

[5] J. W. Rinzler, *The Making of Star Wars Return of the Jedi. The Definitive Story*, Nueva York, Del Rey, 2013, p. 64.

[6] Cass Sunstein, *The World According to Star Wars*, Nueva York, HarperCollins, 2016.

[7] Destacamos aquí el caso simple en el que empieza a obrar un prejuicio. De hecho, incluso en ausencia de tal prejuicio, se puede desarrollar un sesgo hacia una conclusión particular a medida que se acumulan las pruebas debido a la tendencia a la simplicidad y la coherencia. Cuando surge una conclusión provisional, el sesgo de confirmación inclina la recopilación e interpretación de nuevas pruebas a su favor.

[8] Esta observación recibe el nombre de «sesgo de creencia». Véase J. St. B. T. Evans, Julie L. Barson y Paul Pollard, «On the Conflict between Logic and Belief in Syllogistic Reasoning», *Memory & Cognition*, 11, 3 (1983), pp. 295-306.

[9] Dan Ariely, George Loewenstein y Drazen Prelec, «"Coherent Arbitrariness". Stable Demand Curves Without Stable Preferences», *Quarterly Journal of Economics*, 118, 1 (2003), pp. 73-105.

[10] Adam D. Galinsky y T. Mussweiler, «First Offers as Anchors. The Role of Perspective-Taking and Negotiator Focus», *Journal of Personality and Social Psychology*, 81, 4 (2001), pp. 657-669.

[11] Solomon E. Asch, «Forming Impressions of Personality», *Journal of Abnormal and Social Psychology*, 41, 3 (1946), pp. 258-290, fue el primero en usar una serie de adjetivos ordenados de diferente manera para ilustrar este fenómeno.

[12] Steven K. Dallas *et al.*, «Don't Count Calorie Labeling Out. Calorie Counts on the Left Side of Menu Items Lead to Lower Calorie Food Choices», *Journal of Consumer Psychology*, 29, 1 (2019), pp. 60-69.

14. La operación de emparejamiento

[1] S. S. Stevens, «On the Operation Known as Judgment», *American Scientist*, 54, 4 (diciembre de 1966), pp. 385-401. Nuestro uso del término «emparejamiento» es más amplio que el de Stevens, que se limitaba a escalas de *ratio*, de las que trataremos en el capítulo 15.

[2] El ejemplo se utilizó por primera vez en Daniel Kahneman, *Thinking, Fast and Slow*, Nueva York, Farrar, Straus and Giroux, 2011. (Hay trad. cast.: *Pensar rápido, pensar despacio*, Barcelona, Debate, 2012.)

[3] Daniel Kahneman y Amos Tversky, «On the Psychology of Prediction», *Psychological Review*, 80 (1973), pp. 237-251.

[4] G. A. Miller, «The Magical Number Seven, Plus or Minus Two. Some Limits on Our Capacity for Processing Information», *Psychological Review*, 63, 2 (1956), pp. 81-97.

[5] R. D. Goffin y J. M. Olson, «Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others», *Perspectives on Psychological Science*, 6 (2011), pp. 48-60.

15. Escalas

[1] Daniel Kahneman, David Schkade y Cass Sunstein, «Shared Outrage and Erratic Awards. The Psychology of Punitive Damages», *Journal of Risk and Uncertainty*, 16 (1998), pp. 49-86, <https://link. springer. com/ar ticle/10. 1023/A:1007710408413>, y Cass Sunstein, Daniel Kahneman y David Schkade, «Assessing Punitive Damages (with Notes on Cognition and Valuation in Law)», *Yale Law Journal*, 107, 7 (mayo de 1998), pp. 20712153. Los costes de la investigación fueron cubiertos por Exxon en un único acuerdo, pero la compañía no pagó a los investigadores y no tuvo ni control de los datos ni conocimiento previo de los resultados antes de su publicación en revistas académicas.

[2] A. Keane y P. McKeown, *The Modern Law of Evidence*, Nueva York, Oxford University Press, 2014.

[3] Andrew Mauboussin y Michael J. Mauboussin, «If You Say Something Is "Likely", How Likely Do People Think It Is?», *Harvard Business Review*, 3 de julio de 2018.

[4]. BMW *versus* Gore, 517 U.S. 559 (1996), <https://supreme.justia. com/cases/federal/us/517/559>.

[5] Para debates sobre el papel de la emoción en los juicios morales, véanse J. Haidt, «The Emotional Dog and Its Rational Tail. A Social Intuitionist Approach to Moral Judgment», *Psychological Review*, 108, 4 (2001), pp. 814-834, y Joshua Greene, *Moral Tribes. Emotion, Reason, and the Gap Between Us and Them*, Nueva York, Penguin Press, 2014.

[6] Dada la gran cantidad de ruido en estas valoraciones, puede extrañar la altísima correlación (.98) entre los juicios de indignación y de intención punitiva, que proporcionó un respaldo a la hipótesis de la indignación. La extrañeza desaparece cuando se recuerda que la correlación se calculó entre «promedios» de juicios. Para un promedio de cien juicios, el ruido (la desviación estándar de los juicios) se reduce en un factor de 10. El ruido deja de ser un factor cuando se agregan muchos juicios. Véase el capítulo 21.

[7] S. S. Stevens, *Psychophysics. Introduction to Its Perceptual, Neural and Social Prospects*, Nueva York, John Wiley & Sons, 1975.

[8] Dan Ariely, George Loewenstein y Drazen Prelec, «"Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences», *Quarterly Journal of Economics*, 118, 1 (2003), pp. 73-106.

[9] Una transformación en rangos acarrea una pérdida de información, ya que no se conservan las distancias entre los juicios. Supongamos que solo hay tres casos y que un miembro del jurado recomienda una indemnización de diez millones, dos millones y un millón de dólares. Parece claro que el miembro del jurado pretende transmitir una mayor diferencia de intención punitiva entre los dos primeros casos que entre el segundo y el tercero. Sin embargo, una vez convertida en rangos, la diferencia será la misma: una diferencia de un solo rango. Este problema podría resolverse convirtiendo los juicios en puntuaciones estándar.

16. Patrones

[1] R. Blake y N. K. Logothetis, «Visual competition», *Nature Reviews Neuroscience*, 3 (2002), pp. 13-21; M. A. Gernsbacher y M. E. Faust, «The Mechanism of Suppression. A Component of General Comprehension Skill», *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 17 (marzo de 1991), pp. 245-262; M. C. Stites y K. D. Federmeier, «Subsequent to

Suppression. Downstream Comprehension Consequences of Noun/Verb Ambiguity in Natural Reading», *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 41 (septiembre de 2015), pp. 1497-1515.

[2] D. A. Moore y D. Schatz, «The three faces of overconfidence», *Social and Personality Psychology Compass*, 11, 8 (2017), artículo e12331.

[3] S. Highhouse, A. Broadfoot, J. E. Yugo y S. A. Devendorf, «Examining Corporate Reputation Judgments with Generalizability Theory», *Journal of Applied Psychology*, 94 (2009), pp. 782-789. Agradecemos a Scott Highhouse y a Alison Broadfoot por proporcionarnos la información original y a Julian Parris por los análisis complementarios.

[4] P. J. Lamberson y Scott Page, «Optimal forecasting groups», *Management Science*, 58, 4 (2012), pp. 805-810. Agradecemos a Scott Page por atraer nuestra atención sobre la fuente de este ruido de patrón.

[5] El trabajo de Allport y Odbert (1936) sobre el vocabulario inglés relevante para la personalidad se cita en Oliver P. John y Sanjay Strivastava, «The Big-Five Trait Taxonomy. History, Measurement, and Theoretical Perspectives», en *Handbook of Personality. Theory and Research*, 2.ª ed., L. Pervin y Oliver P. John (eds.), Nueva York, Guilford, 1999.

[6] Ian W. Eisenberg, Patrick G. Bissett, A. Zeynep Enkavi *et al*., «Uncovering the structure of self-regulation through data-driven ontology discovery», *Nature Communications*, 10 (2019), p. 2319.

[7] Walter Mischel, «Toward an integrative science of the person», *Annual Review of Psychology*, 55 (2004), pp. 1-22.

17. Las fuentes de ruido

[1] Aunque no existe una regla general sobre el desglose del sesgo y el ruido, las proporciones de esta figura son más o menos representativas de algunos de los ejemplos, reales o ficticios, que hemos revisado. En concreto, en esta figura, el sesgo y el ruido son iguales (como lo eran en las predicciones de ventas de GoodSell). El cuadrado del ruido de nivel representa el 37 por ciento del cuadrado del ruido del sistema (como ocurrió en el estudio de las indemnizaciones punitivas). El cuadrado del ruido de ocasión es, como se muestra, aproximadamente, el 35 por ciento del cuadrado del ruido de patrón.

[2] Véanse las referencias en la introducción. Mark A. Lemley y Bhaven Sampat, «Examiner Characteristics and Patent Office Outcomes», *Review of Economics and Statistics*, 94, 3 (2012), pp. 817-827. Véanse también Iain Cockburn, Samuel Kortum y Scott Stern, «Are All Patent Examiners Equal? The Impact of Examiner Characteristics», documento de trabajo 8980, junio de 2002, <www.nber.org/papers/w8980>, y Michael D. Frakes y Melissa F. Wasserman, «Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data», *Review of Economics and Statistics*, 99, 3 (julio de 2017), pp. 550-563.

[3] Joseph J. Doyle Jr., «Child Protection and Child Outcomes. Measuring the Effects of Foster Care», *American Economic Review*, 95, 5 (diciembre de 2007), pp. 1583-1610.

[4] Andrew I. Schoenholtz, Jaya Ramji-Nogales y Philip G. Schrag, «Refugee Roulette. Disparities in Asylum Adjudication», *Stanford Law Review*, 60, 2 (2007).

[5] Este valor se estima a partir de los cálculos presentados en el capítulo 6, donde la varianza de la interacción es el 23 por ciento de la varianza total. Suponiendo que las sentencias se distribuyen

normalmente, la diferencia media absoluta entre dos observaciones seleccionadas al azar es de 1,128 DE.

[6] J. E. Martinez, B. Labbree, S. Uddenberg y A. Todorov, «Meaningful "noise": Comparative judgments contain stable idiosyncratic contributions» (ms. no publicado).

[7] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig y S. Mullainathan, «Human Decisions and Machine Predictions», *Quarterly Journal of Economics*, 133 (2018), pp. 237-293.

[8] El modelo produjo para cada juez tanto una ordenación de los 141.833 casos como un umbral a partir del cual se concedería la libertad bajo fianza. El ruido de nivel refleja la variabilidad de los umbrales, mientras que el ruido de patrón muestra la variabilidad en la ordenación de los casos.

[9] Gregory Stoddard, Jens Ludwig y Sendhil Mullainathan, intercambios de correos electrónicos con los autores en junio-julio de 2020.

18. Mejores juzgadores para mejores juicios

[10] Phil Rosenzweig, *Left brain, right stuff. How leaders make winning decisions*, Nueva York, Public Affairs, 2014.

[1] Albert E. Mannes *et al.*, «The Wisdom of Select Crowds», *Journal of Personality and Social Psychology*, 107, 2 (2014), pp. 276-299; Jason Dana *et al.*, «The Composition of Optimally Wise Crowds», *Decision Analysis*, 12, 3 (2015), pp. 130-143.

[2] Briony D. Pulford, Andrew M. Colmna, Eike K. Buabang y Eva M. Krockow, «The Persuasive Power of Knowledge. Testing the Confidence Heuristic», *Journal of Experimental Psychology. General*, 147, 10 (2018), pp. 1431-1444.

[3] Nathan R. Kuncel y Sarah A. Hezlett, «Fact and Fiction in Cognitive Ability Testing for Admissions and Hiring Decisions», *Current Directions in Psychological Science*, 19, 6 (2010), pp. 339-345.

[4]. *Ibid*.

[5] Frank L. Schmidt y John Hunter, «General Mental Ability in the World of Work. Occupational Attainment and Job Performance», *Journal of Personality and Social Psychology*, 86, 1 (2004), pp. 162-173.

[6] Angela L. Duckworth, David Weir, Eli Tsukayama y David Kwok, «Who Does Well in Life? Conscientious Adults Excel in Both Objective and Subjective Success», *Frontiers in Psychology*, 3 (septiembre de 2012). Sobre el término *grit*, véase Angela L. Duckworth, Christopher Peterson, Michael D. Matthews y Dennis Kelly, «Grit. Perseverance and Passion for Long-Term Goals», *Journal of Personality and Social Psychology*, 92, 6 (2007), pp. 1087-1101.

[7] Richard E. Nisbett *et al.*, «Intelligence. New Findings and Theoretical Developments», *American Psychologist*, 67, 2 (2012), pp. 130-159.

[8] F. L. Schmidt y J. Hunter, «General Mental Ability...», p. 162.

[9] N. R. Kuncel y S. A. Hezlett, «Fact and Fiction...».

[10] Estas correlaciones se derivan de metanálisis que corrigen las correlaciones observadas por error de medición en el criterio y la restricción de rango. Existen debates entre los investigadores sobre si estas correcciones exageran el valor predictivo de la CIG. Sin embargo, dado que estos debates metodológicos se aplican también a otros predictores, los expertos suelen estar de acuerdo en que la CIG (junto con los test de muestras de trabajo [véase el capítulo 24]) es el mejor predictor disponible del éxito laboral. Véase N. R. Kuncel y S. A. Hezlett, «Fact and Fiction...».

[11] F. L. Schmidt y J. Hunter, «General Mental Ability...», p. 162.

[12] David Lubinski, «Exceptional Cognitive Ability. The Phenotype», *Behavior Genetics*, 39, 4 (2009), pp. 350-358.

[13] Jonathan Wai, «Investigating America's Elite. Cognitive Ability, Education, and Sex Differences», *Intelligence*, 41, 4 (2013), pp. 203-211.

[14] Keela S. Thomson y Daniel M. Oppenheimer, «Investigating an Alternate Form of the Cognitive Reflection Test», *Judgment and Decision Making*, 11, 1 (2016), pp. 99-113.

[15] Gordon Pennycook *et al*., «Everyday Consequences of Analytic Thinking», *Current Directions in Psychological Science*, 24, 6 (2015), pp. 425432.

[16] Gordon Pennycook y David G. Rand, «Lazy, Not Biased. Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning», *Cognition*, 188 (junio de 2018), pp. 39-50.

[17] Nathaniel Barr *et al*., «The Brain in Your Pocket. Evidence That Smartphones Are Used to Supplant Thinking», *Computers in Human Behavior*, 48 (2015), pp. 473-480.

[18] Niraj Patel, S. Glenn Baker y Laura D. Scherer, «Evaluating the Cognitive Reflection Test as a Measure of Intuition/Reflection, Numeracy, and Insight Problem Solving, and the Implications for Understanding Real-World Judgments and Beliefs», *Journal of Experimental Psychology. General*, 148, 12 (2019), pp. 2129-2153.

[19] John T. Cacioppo y Richard E. Petty, «The Need for Cognition», *Journal of Personality and Social Psychology*, 42, 1 (1982), pp. 116-131.

[20] Stephen M. Smith e Irwin P. Levin, «Need for Cognition and Choice Framing Effects», *Journal of Behavioral Decision Making*, 9, 4 (1996), pp. 283-290.

[21] Judith E. Rosenbaum y Benjamin K. Johnson, «Who's Afraid of Spoilers? Need for Cognition, Need for Affect, and Narrative Selection and Enjoyment», *Psychology of Popular Media Culture*, 5, 3 (2016), pp. 273-289.

[22] Wandi Bruine De Bruin *et al*., «Individual Differences in Adult Decision-Making Competence», *Journal of Personality and Social Psychology*, 92, 5 (2007), pp. 938-956.

[23] Heather A. Butler, «Halpern Critical Thinking Assessment Predicts Real-World Outcomes of Critical Thinking», *Applied Cognitive Psychology*, 26, 5 (2012), pp. 721-729.

[24] Uriel Haran, Ilana Ritov y Barbara Mellers, «The Role of Actively Open-Minded Thinking in Information Acquisition, Accuracy, and Calibration», *Judgment and Decision Making*, 8, 3 (2013), pp. 188-201.

[25] *Ibid*.

19. Supresión de sesgos e higiene de las decisiones

[26] J. Baron, «Why Teach Thinking? An Essay», *Applied Psychology. An International Review*, 42 (1993), pp. 191-214; J. Baron, *The Teaching of Thinking. Thinking and Deciding*, 2.ª ed., Nueva York, Cambridge University Press, 1994, pp. 127-148.

[1] Una excelente exposición es la de Jack B. Soll *et al*., «A User's Guide to Debiasing», en *The Wiley Blackwell Handbook of Judgment and Decision Making*, Gideon Keren y George Wu (eds.), vol. 2, Nueva York, John Wiley & Sons, 2015, p. 684.

[2] HM Treasury, *The Green Book. Central Government Guidance on Appraisal and Evaluation*, Londres, 2018, <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/ 685903/The_Green_Book.pdf>.

[3] Richard H. Thaler y Cass R. Sunstein, *Nudge. Improving Decisions about Health, Wealth, and Happiness*, New Haven, CT, Yale University Press, 2008.

[4] Ralph Hertwig y Till Grüne-Yanoff, «Nudging and Boosting. Steering or Empowering Good Decisions», *Perspectives on Psychological Science*, 12, 6 (2017).

[5] Geoffrey T. Fong *et al*., «The Effects of Statistical Training on Thinking About Everyday Problems», *Cognitive Psychology*, 18, 3 (1986), pp. 253292.

[6] Willem A. Wagenaar y Gideon B. Keren, «Does the Expert Know? The Reliability of Predictions and Confidence Ratings of Experts», en *Intelligent Decision Support in Process Environments*, Erik Hollnagel, Giuseppe Mancini y David D. Woods (eds.), Berlín, Springer, 1986, pp. 87-103.

[7] Carey K. Morewedge *et al*., «Debiasing Decisions. Improved Decision Making with a Single Training Intervention», *Policy Insights from the Behavioral and Brain Sciences*, 2, 1 (2015), pp. 129-140.

[8] Anne-Laure Sellier *et al*., «Debiasing Training Transfers to Improve Decision Making in the Field», *Psychological Science*, 30, 9 (2019), pp. 13711379.

[9] Emily Pronin *et al*.,«The Bias Blind Spot. Perceptions of Bias in Self Versus Others», *Personality and Social Psychology Bulletin*, 28, 3 (2002), pp. 369381.

[10] Daniel Kahneman, Dan Lovallo y Olivier Sibony, «Before You Make That Big Decision...», *Harvard Business Review*, 89, 6 (junio de 2011), pp. 50-60.

[11] Atul Gawande, *Checklist Manifesto. How to Get Things Right*, Nueva York, Metropolitan Books, 2010.

[12] Office of Information and Regulatory Affairs, «Agency Checklist. Regulatory Impact Analysis», sin fecha, <www.whitehouse.gov/sites/whitehouse.gov/files/omb/inforeg/inforeg/regpol/RIA_Checklist.pdf>.

[13] Esta lista de control ha sido parcialmente adaptada de Daniel Kahneman *et al*., «Before You Make That Big Decision...».

20. Secuenciación de la información en la ciencia forense

[14] Véase A. Gawande, *Checklist Manifesto*...

[1] R. Stacey, «A Report on the Erroneous Fingerprint Individualisation in the Madrid Train Bombing Case», *Journal of Forensic Identification*, 54 (2004), pp. 707-718.

[2] Michael Specter, «Do Fingerprints Lie?», *The New Yorker*, 27 de mayo de 2002. La cursiva es nuestra.

[3] I. E. Dror y R. Rosenthal, «Meta-analytically Quantifying the Reliability and Biasability of Forensic Experts», *Journal of Forensic Science*, 53 (2008), pp. 900-903.

[4] I. E. Dror, D. Charlton y A. E. Péron, «Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications», *Forensic Science International*, 156 (2006), pp. 74-78.

[5] I. E. Dror y D. Charlton, «Why Experts Make Errors», *Journal of Forensic Identification*, 56 (2006), pp. 600-616.

[6] I. E. Dror y S. A. Cole, «The Vision in “Blind” Justice. Expert Perception, Judgment, and Visual Cognition in Forensic Pattern Recognition», *Psychonomic Bulletin and Review*, 17 (2010), pp. 161-167, 165. Véase también I. E. Dror, «A Hierarchy of Expert Performance (HEP)», *Journal of Applied Research in Memory and Cognition*, 5 (2016), pp. 121-127.

[7] I. E. Dror *et al*., «Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a "Target" Comparison», *Forensic Science International*, 208 (2011), pp. 10-17.

[8] B. T. Ulery, R.A.Hicklin, M. A. Roberts y J.A.Buscaglia,«Changes in Latent Fingerprint Examiners' Markup Between Analysis and Comparison», *Forensic Science International*, 247 (2015), pp. 54-61.

[9] I. E. Dror y G. Hampikian, «Subjectivity and Bias in Forensic DNA Mixture Interpretation», *Science and Justice*, 51 (2011), pp. 204-208.

[10] M. J.Saks, D.M. Risinger, R. Rosenthal y W. C. Thompson, «Context Effects in Forensic Science: a Review and Application of the Science of Science to Crime Laboratory Practice in the United States», *Science Justice Journal of Forensic Science Society*, 43 (2003), pp. 77-90.

[11] PCAST (Gabinete de asesores en ciencia y tecnología del presidente), *Report to the President. Forensic Science in Criminal Courts. Ensuring Scientific Validity of Feature-Comparison Methods*, Washington D. C., Executive Office of the President, PCAST, 2016.

[12] R. Stacey, «A Report on the Erroneous Fingerprint...».

[13] I. E. Dror y S. A. Cole, «The Vision in "Blind" Justice...».

[14] I. E. Dror, «Biases in Forensic Experts», *Science*, 360 (2018), p. 243.

[15] I. E. Dror y D. Charlton, «Why Experts Make Errors».

[16] B. T. Ulery, R. A. Hicklin, J. A. Buscaglia y M. A. Roberts, «Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners», *PLoS One*, 7 (2012).

[17] Innocence Project, «Overturning Wrongful Convictions Involving Misapplied Forensics», *Misapplication of Forensic Science*, (2018), pp. 1-7, <www.innocenceproject.org/causes/misapplication-forensic-science>. Véase también S. M. Kassin, I. E. Dror, J. Kukucka y L. Butt, «The Forensic Confirmation Bias. Problems, Perspectives, and Proposed Solutions», *Journal of Applied Research in Memory and Cognition*, 2 (2013), pp. 42-52.

[18] PCAST, *Report to the President*.

[19] B. T. Ulery, R. A. Hicklin, J. Buscaglia y M. A. Roberts, «Accuracy and Reliability of Forensic Latent Fingerprint Decisions», *Proceedings of the National Academy of Sciences*, 108 (2011), pp. 7733-7738.

[20] PCAST, *Report to the President*, p. 95. En cursiva en el original.

[21] Igor Pacheco, Brian Cerchiai y Stephanie Stoiloff, «Miami-Dade Research Study for the Reliability of the ACE-V Process. Accuracy & Precision in Latent Fingerprint Examinations», informe definitivo, Miami-Dade Police Department Forensic Services Bureau, 2014, <www.ncjrs.gov/ pdffiles1/nij/grants/248534.pdf>.

[22] B.T.Ulery, R.A.Hicklin, M.A.Roberts y J.A.Buscaglia,«Factors Associated with Latent Fingerprint Exclusion Determinations», *Forensic Science International*, 275 (2017), pp. 65-75.

[23] R. N. Haber e I. Haber, «Experimental Results of Fingerprint Comparison Validity and Reliability. A Review and Critical Analysis», *Sci. Justice*, 54 (2014), pp. 375-389.

[24] I. E. Dror, «A Hierarchy of Expert Performance (HEP)», p. 3.

[25] M. Leadbetter, carta al director, *Fingerprint World*, 33 (2007), p. 231.

[26] L. Butt, «The Forensic Confirmation Bias. Problems, Perspectives, and Proposed Solutions – Commentary by a Forensic Examiner», *Journal of Applied Research in Memory and Cognition*, 2 (2013), pp. 59-60. Las cursivas son nuestras.

[27] R. Stacey, «A Report on the Erroneous Fingerprint...», p. 713. Las cursivas son nuestras.

[28] J. Kukucka, S. M. Kassin, P. A. Zapf e I. E. Dror, «Cognitive Bias and Blindness. A Global Survey of Forensic Science Examiners», *Journal of Applied Research in Memory and Cognition*, 6 (2017).

[29] I. E. Dror *et al*., carta al director, «Context Management Toolbox. A Linear Sequential Unmasking (LSU) Approach for Minimizing Cognitive Bias in Forensic Decision Making», *Journal of Forensic Science*, 60 (2015), pp. 1111-1112.

[1] Jeffrey A. Frankel, «Over-optimism in Forecasts by Official Budget Agencies and Its Implications», documento de trabajo 17239, National Bureau of Economic Research, diciembre de 2011, <www.nber.org/papers/w17239>.

[2] H. R. Arkes, «Overconfidence in Judgmental Forecasting», en *Principles of Forecasting. A Handbook for Researchers and Practitioners*, Jon Scott Armstrong (ed.), vol. 30, International Series in Operations Research & Management Science, Boston, Springer, 2001.

[3] Itzhak Ben-David, John Graham y Campell Harvey, «Managerial Miscalibration», *The Quarterly Journal of Economics*, 128, 4 (noviembre de 2013), pp. 1547-1584.

[4] T. R. Stewart, «Improving Reliability of Judgmental Forecasts», en *Principles of Forecasting. A Handbook for Researchers and Practitioners*, Jon Scott Armstrong (ed.), vol. 30, International Series in Operations Research & Management Science, Boston, Springer, 2001 (en adelante citado como *Principles of Forecasting*...), p. 82.

[5] Theodore W. Ruger, Pauline T. Kim, Andrew D. Martin y Kevin M. Quinn, «The Supreme Court Forecasting Project. Legal and Political Science Approaches to Predicting Supreme Court Decision-Making», *Columbia Law Review*, 104 (2004), pp. 1150-1209.

[6] Cass Sunstein, «Maximin», *Yale Journal of Regulation* (borrador; 3 de mayo de 2020), <https://papers.ssrn.com/sol3/papers.cfm?abstract_ id=3476250>.

[7] Para numerosos ejemplos, véase J. S. Armstrong (ed.), *Principles of Forecasting*...

[8] Jon Scott Armstrong, «Combining Forecasts», en *Principles of Forecasting*..., pp. 417-439.

[9] T. R. Stewart, «Improving Reliability of Judgmental Forecasts», en *Principles of Forecasting*..., p. 95.

[10] J. S. Armstrong, «Combining Forecasts».

[11] Albert E. Mannes *et al*., «The Wisdom of Select Crowds», *Journal of Personality and Social Psychology*, 107, 2 (2014), pp. 276-299.

[12] Justin Wolfers y Eric Zitzewitz, «Prediction Markets», *Journal of Economic Perspectives*, 18 (2004), pp. 107-126.

[13] Cass R. Sunstein y Reid Hastie, *Wiser. Getting Beyond Groupthink to Make Groups Smarter*, Boston, Harvard Business Review Press, 2014.

[14] Gene Rowe y George Wright, «The Delphi Technique as a Forecasting Tool. Issues and Analysis», *International Journal of Forecasting*, 15 (1999), pp. 353-375. Véase también Dan Bang y Chris D. Frith, «Making Better Decisions in Groups», *Royal Society Open Science*, 4, 8 (2017).

[15] R. Hastie, «Review essay. Experimental evidence on group accuracy», en B. Grofman y G. Guillermo (eds.), *Information Pooling and Group Decision Making*, Greenwich, Connecticut, JAI Press, 1986, pp. 129-157.

[16] Andrew H. Van de Ven y André L. Delbecq, «The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes», *Academy of Management Journal*, 17, 4 (2017).

[17] Philip E. Tetlock y Dan Gardner, *Superforecasting. The Art and Science of Prediction*, Nueva York, Random House, 2016, p. 95.

[18] *Ibid*., p. 231.

[19] *Ibid*., p. 273.

[20] Ville A. Satopää, Marat Salikhov, Philip E. Tetlock y Barb Mellers, «Bias, Information, Noise. The BIN Model of Forecasting», 19 de febrero de 2020, 23, <https://dx.doi.org/10.2139/ssrn.3540864>.

[21] *Ibid*., p. 23.

[22] *Ibid*., p. 22.

[23] *Ibid*., p. 24.

22. Directrices en la medicina

[24] Clintin P. Davis-Stober, David V. Budescu, Stephen B. Broomell y Jason Dana, «The Composition of Optimally Wise Crowds», *Decision Analysis*, 12, 3 (2015), pp. 130-143.

[1] Laura Horton *et al*., «Development and Assessment of Inter- and Intra-Rater Reliability of a Novel Ultrasound Tool for Scoring Tendon and Sheath Disease: A Pilot Study», *Ultrasound*, 24, 3 (2016), p. 134, <www. ncbi.nlm.nih.gov/pmc/articles/PMC5105362>.

[2] Laura C. Collins *et al*., «Diagnostic Agreement in the Evaluation of Image-guided Breast Core Needle Biopsies», *American Journal of Surgical Pathology*, 28 (2004), p. 28, <https://journals.lww.com/ajsp/Abstract/2004/01000/Diagnostic_Agreement_in_the_Evaluation_of.15.aspx>.

[3] Julie L. Fierro *et al*., «Variability in the Diagnosis and Treatment of Group A Streptococcal Pharyngitis by Primary Care Pediatricians», *Infection Control and Hospital Epidemiology*, 35, S3 (2014), S79, <www.jstor.org/sta ble/10.1086/677820>.

[4] DiabetesTests, Centers for Disease, Control and Prevention, <https://www.cdc.gov/diabetes/basics/getting-tested.html> (consultado el 15 de enero de 2020).

[5] Joseph D. Kronz *et al*., «Mandatory Second Opinion Surgical Pathology at a Large Referral Hospital», *Cancer*, 86 (1999), p. 2426, <https://onlinelibrary.wiley.com/doi/full/10.1002/(SICI)1097-0142(19991201)86:11%3C2426::AID-CNCR34%3E3.0.CO;2-3>.

[6] La mayor parte del material puede encontrarse en línea; un resumen en libro es Dartmouth Medical School, *The Quality of Medical Care in the United States. A Report on the Medicare Program; the Dartmouth Atlas of Health Care 1999*, American Hospital Publishers, 1999.

[7] Véase, por ejemplo, OECD, *Geographic Variations in Health Care: What Do We Know and What Can Be Done to Improve Health System Performance*, París, OECD Publishing, 2014, pp. 137-169; Michael P. Hurley *et al*., «Geographic Variation in Surgical Outcomes and Cost Between the United States and Japan», *American Journal of Managed Care*, 22 (2016), p. 600, <www.ajmc.com/journals/issue/2016/2016-vol22-n9/geographic-varia tion-in-surgical-outcomes-and-cost-between-the-united-states-and-ja pan>; John Appleby, Veena Raleigh, Francesca Frosini, Gwyn Bevan, Haiyan Gao y Tom Lyscom, *Variations in Health Care: The Good, the Bad and the Inexplicable*, Londres, The King's Fund, 2011, <www.kingsfund.org.uk/sites/default/files/Variations-in-health-care-good-bad-inexplicable-reportThe-Kings-Fund-April-2011.pdf>.

[8] David C. Chan Jr. *et al*., «Selection with Variation in Diagnostic Skill. Evidence from Radiologists», documento de trabajo 26467, National Bureau of Economic Research, noviembre de 2019, <www.nber.org/pa pers/w26467>.

[9] P. J. Robinson, «Radiology's Achilles' Heel. Error and Variation in the Interpretation of the Röntgen Image», *British Journal of Radiology*, 70 (1997), p. 1085,

<www.ncbi.nlm.nih.gov/pubmed/9536897>. Un estudio relevante es el de Yusuke Tsugawa *et al*., «Physician Age and Outcomes in Elderly Patients in Hospital in the US. Observational Study», *BMJ*, 357 (2017), <www.bmj.com/content/357/bmj.j1797>, que encuentra que los resultados de los médicos empeoran cuanto más lejos están de formarse. De ello se sigue que existe un equilibrio entre el desarrollo de la experiencia, resultante de los años de práctica, y la familiaridad con las pruebas y las directrices más recientes. El estudio concluye que los mejores resultados proceden de los médicos que se encuentran en los primeros años de residencia, cuando tienen esas pruebas en mente.

[10] P. J. Robinson, «Radiology's Achilles' Heel...».

[11] Al igual que el coeficiente de correlación, la kappa puede ser negativa, aunque es algo raro en la práctica. He aquí una caracterización del significado de las diferentes estadísticas kappa: «leve (κ = 0,00 - 0,20), regular (κ = 0,21 - 0,40), moderado (κ = 0,41 - 0,60), sustancial (κ = 0,61 - 0,80) y casi total (κ > 0,80)», en Ron Wald, Chaim M. Bell, Rosane Nisenbaum, Samuel Perrone, Orfeas Liangos, Andreas Laupacis y Bertrand L. Jaber, «Interobserver Reliability of Urine Sediment Interpretation», *Clinical Journal of the American Society of Nephrology*, 4, 3 (marzo de 2009), pp. 567-571, <https: //cjasn.asnjournals.org/content/4/3/567>.

[12] Howard R. Strasberg *et al*., «Inter-Rater Agreement Among Physicians on the Clinical Significance of Drug-Drug Interactions», *AMIA Annual Symposium Proceedings*, noviembre de 2013, p. 1325, <www.ncbi. nlm.nih.gov/pmc/articles/PMC3900147>.

[13] Wald *et al*., «Interobserver Reliability of Urine Sediment Interpretation», <https://cjasn.asnjournals.org/content/4/3/567>.

[14] Juan P. Palazzo *et al*., «Hyperplastic Ductal and Lobular Lesions and Carcinomas in Situ of the Breast. Reproducibility of Current Diagnostic Criteria Among Community- and Academic-Based Pathologists», *Breast Journal*, 4 (2003), p. 230, <www.ncbi.nlm.nih.gov/pubmed/21223441>.

[15] Rohit K. Jain *et al*., «Atypical Ductal Hyperplasia. Interobserver and Intraobserver Variability», *Modern Pathology*, 24 (2011), p. 917, <www. nature.com/articles/modpathol201166>.

[16] Alex C. Speciale *et al*., «Observer Variability in Assessing Lumbar Spinal Stenosis Severity on Magnetic Resonance Imaging and Its Relation to Cross-Sectional Spinal Canal Area», *Spine*, 27 (2002), p. 1082, <www. ncbi.nlm.nih.gov/pubmed/12004176>.

[17] Centers for Disease Control and Prevention, «Heart Disease Facts», consultado el 16 de junio de 2020, <www.cdc.gov/heartdisease/facts. htm>.

[18] Timothy A. DeRouen *et al*., «Variability in the Analysis of Coronary Arteriograms», *Circulation*, 55 (1977), p. 324, <www.ncbi.nlm.nih. gov/pubmed/832349>.

[19] Olaf Buchweltz *et al*., «Interobserver Variability in the Diagnosis of Minimal and Mild Endometriosis», *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 122 (2005), p. 213, <www.ejog.org/article/ S0301-2115(05)00059-X/pdf>.

[20] Jean-Pierre Zellweger *et al*., «Intra-observer and Overall Agreement in the Radiological Assessment of Tuberculosis», *International Journal of Tuberculosis & Lung Disease*, 10 (2006), p. 1123, <www.ncbi.nlm.nih.gov/ pubmed/17044205>. Para la concordancia «regular» entre evaluadores, véase Yanina Balabanova *et al*., «Variability in Interpretation of Chest Radiographs Among Russian Clinicians and Implications for Screening Programmes. Observational Study», *BMJ*, 331 (2005), p. 379, <www.bmj.com/content/ 331/7513/379.short>.

[21] Shinsaku Sakurada *et al*., «Inter-Rater Agreement in the Assessment of Abnormal Chest X-Ray Findings for Tuberculosis Between Two Asian Countries», *BMC Infectious Diseases*, 12, 31 (2012), <https://bmcin fectdis.biomedcentral.com/articles/10.1186/1471-2334-12-31>.

[22] Evan R. Farmer *et al*., «Discordance in the Histopathologic Diagnosis of Melanoma and Melanocytic Nevi Between Expert Pathologists», *Human Pathology*, 27 (1996), p. 528,

<www.ncbi.nlm.nih.gov/pubmed/ 8666360>.

[23] Alfred W. Kopf, M. Mintzis y R. W. Bart, «Diagnostic Accuracy in Malignant Melanoma», *Archives of Dermatology*, 111 (1975), p. 1291, <www. ncbi.nlm.nih.gov/pubmed/1190800>.

[24] Maria Miller y A. Bernard Ackerman, «How Accurate Are Dermatologists in the Diagnosis of Melanoma? Degree of Accuracy and Implications», *Archives of Dermatology*, 128 (1992), p. 559, <https://jamanetwork. com/journals/jamadermatology/fullarticle/554024>.

[25] Craig A. Beam *et al*., «Variability in the Interpretation of Screening Mammograms by US Radiologists», *Archives of Internal Medicine*, 156 (1996), p. 209, <www.ncbi.nlm.nih.gov/pubmed/8546556>.

[26] P. J. Robinson *et al*., «Variation Between Experienced Observers in the Interpretation of Accident and Emergency Radiographs», *British Journal of Radiology*, 72 (1999), p. 323, <www.birpublications.org/doi/pdf/10.1259/ bjr.72.856.10474490>.

[27] Katherine M. Detre *et al*., «Observer Agreement in Evaluating Coronary Angiograms», *Circulation*, 52 (1975), p. 979, <www.ncbi.nlm.nih. gov/pubmed/1102142>.

[28] L. Horton *et al*., «Development and Assessment...»; Megan Banky *et al*., «Inter- and Intra-Rater Variability of Testing Velocity When Assessing Lower Limb Spasticity», *Journal of Rehabilitation Medicine*, 51 (2018), <www. medicaljournals.se/jrm/content/abstract/10.2340/16501977-2496>.

[29] Esther Y. Hsiang *et al*., «Association of Primary Care Clinic Appointment Time with Clinician Ordering and Patient Completion of Breast and Colorectal Cancer Screening», *JAMA Network Open*, 51 (2019), <https://ja manetwork.com/journals/jamanetworkopen/fullarticle/2733171>.

[30] Hengchen Dai *et al*., «The Impact of Time at Work and Time Offfrom Work on Rule Compliance. The Case of Hand Hygiene in Health Care», *Journal of Applied Psychology*, 100 (2015), p. 846, <www.ncbi.nlm.nih. gov/pubmed/25365728>.

[31] Ali S. Raja, «The HEART Score Has Substantial Interrater Reliability», *NEJM J Watch*, 5 de diciembre de 2018, <www.jwatch.org/ na47998/2018/12/05/heart-score-has-substantial-interrater-reliability> (reseña en Colin A. Gershon *et al*., «Inter-rater Reliability of the HEART Score», *Academic Emergency Medicine*, 26 [2019], p. 552).

[32] Jean-Pierre Zellweger *et al*., «Intra-observer and Overall Agreement in the Radiological Assessment of Tuberculosis», *International Journal of Tuberculosis & Lung Disease*, 10 (2006), p. 1123, <www.ncbi.nlm.nih.gov/ pubmed/17044205>; Ibrahim Abubakar *et al*., «Diagnostic Accuracy of Digital Chest Radiography for Pulmonary Tuberculosis in a UK Urban Population», *European Respiratory Journal*, 35 (2010), p. 689, <https://erj. ersjournals.com/content/35/3/689.short>.

[33] Michael Barnett *et al*., «Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians», *JAMA Network Open*, 2 (2019), e19009, <https://jamanetwork.com/jour nals/jamanetworkopen/fullarticle/2726709>; Kimberly H. Allison *et al*., «Understanding Diagnostic Variability in Breast Pathology. Lessons Learned from an Expert Consensus Review Panel», *Histopathology*, 65 (2014), p. 240, <https://onlinelibrary.wiley.com/doi/abs/10.1111/his.12387>.

[34] Babak Ehteshami Bejnordi *et al*., «Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer», *JAMA*, 318 (2017), p. 2199, <https://jamanetwork. com/journals/jama/fullarticle/2665774>.

[35] Varun Gulshan *et al*., «Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs», *JAMA*, 316 (2016), p. 2402, <https://jamanetwork.com/ journals/jama/fullarticle/2588763>.

[36] Mary Beth Massat, «A Promising Future for AI in Breast Cancer Screening», *Applied Radiology*, 47 (2018), p. 22, <www.appliedradiology.com/ articles/a-promising-future-for-ai-in-breast-cancer-screening>; Alejandro Rodriguez-Ruiz *et al*., «Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography. Comparison with 101 Radiologists», *Journal of the National Cancer Institute*, 111 (2019), p. 916, <https://academic.oup.com/ jnci/advance-article-abstract/doi/10.1093/jnci/djy222/5307077>.

[37] Puntuación de Apgar, Medline Plus, <https://medlineplus.gov/ ency/article/003402.htm> (consultado el 4 de febrero de 2020).

[38] L. R. Foster *et al*., «The Interrater Reliability of Apgar Scoresat 1 and 5 Minutes», *Journal of Investigative Medicine*, 54, 1 (2006), p. 293, <https://jim.bmj.com/content/54/1/S308.4>.

[39] Warren J. McIsaac *et al*., «Empirical Validation of Guidelines for the Management of Pharyngitis in Children and Adults», *JAMA*, 291 (2004), p. 1587, <www.ncbi.nlm.nih.gov/pubmed/15069046>.

[40] Emilie A.Ooms *et al*., «Mammography. Interobserver Variability in Breast Density Assessment», *Breast*, 16 (2007), p. 568, <www.sciencedirect. com/science/article/abs/pii/S0960977607000793>.

[41] Frances P. O'Malley et al., «Interobserver Reproducibility in the Diagnosis of Flat Epithelial Atypia of the Breast», *Modern Pathology*, 19 (2006), p. 172, <www.nature.com/articles/3800514>.

[42] Véase Ahmed Aboraya *et al*., «The Reliability of Psychiatric Diagnosis Revisited», *Psychiatry (Edgmont)*, 3 (2006), p. 41, <www.ncbi.nlm.nih. gov/pmc/articles/PMC2990547>. Para un panorama general, véase N. Kreitman, «The Reliability of Psychiatric Diagnosis», *Journal of Mental Science*, 107 (1962), pp. 876-886, <www.cambridge.org/core/journals/journalof-mental-science/article/reliability-of-psychiatric-diagnosis/92832FFA 170F4FF41189428C6A3E6394>.

[43] A. Aboraya *et al*., «Reliability of Psychiatric Diagnosis Revisited», p. 43.

[44] C. H. Ward *et al*., «The Psychiatric Nomenclature. Reasons for Diagnostic Disagreement», *Archives of General Psychiatry*, 7 (1962), p. 198.

[45] A. Aboraya *et al*., «Reliability of Psychiatric Diagnosis Revisited».

[46] Samuel M. Lieblich, David J. Castle, Christos Pantelis, Malcolm Hopwood, Allan Hunter Young e Ian P. Everall, «High Heterogeneity and Low Reliability in the Diagnosis of Major Depression Will Impair the Development of New Drugs», *British Journal of Psychiatry Open*, 1 (2015), e5e7, <www.ncbi.nlm.nih.gov/pmc/articles/PMC5000492/pdf/bjpor cpsych_1_2_e5.pdf>.

[47] *Ibid*.

[48]. Véanse Elie Cheniaux *et al*., «The Diagnoses of Schizophrenia, Schizoaffective Disorder, Bipolar Disorder and Unipolar Depression. Interrater Reliability and Congruence Between DSM-IV and ICD-10», *Psychopathology*, 42 (2009), pp. 296-298, en particular p. 293, y Michael Chmielewski *et al*., «Method Matters. Understanding Diagnostic Reliability in DSM-IV and DSM-5», *Journal of Abnormal Psychology*, 124 (2015), pp. 764, 768-769.

[49] A. Aboraya *et al*., «Reliability of Psychiatric Diagnosis Revisited», p. 47.

[50] *Ibid*., p. 47.

[51] Véase M. Chmielewski *et al*., «Method Matters...».

[52] Véase, por ejemplo, Helena Chmura Kraemer *et al*., «DSM-5. How Reliable Is Reliable Enough?», *American Journal of Psychiatry*, 169 (2012), pp. 13-15.

[53] S. M. Lieblich *et al*., «High Heterogeneity...».

[54] *Ibid*., e-5.

[55] *Ibid*.

[56] *Ibid*., e-6.
[57] A. Aboraya *et al*., «Reliability of Psychiatric Diagnosis Revisited», p. 47.
[58] *Ibid*.
[59] *Ibid*.

23. Definición de la escala en las calificaciones

[60] Se pueden encontrar algunas valiosas notas de advertencia en Christopher Worsham y Anupam B. Jena, «The Art of Evidence-Based Medicine», *Harvard Business Review*, 30 de enero de 2019, <https://hbr. org/2019/01/the-art-of-evidence-based-medicine>.

[1] Jena McGregor, «Study Finds That Basically Every Single Person Hates Performance Reviews», *The Washington Post*, 27 de enero de 2014.

[2] La transformación digital que están experimentando muchas organizaciones puede crear aquí nuevas posibilidades. En teoría, las empresas pueden ahora recopilar grandes cantidades de pequeña información en tiempo real sobre el rendimiento de cada trabajador. Estos datos pueden hacer posibles evaluaciones del rendimiento totalmente algorítmicas para algunos puestos. Sin embargo, aquí nos centramos en los puestos en los que el juicio no puede eliminarse por completo de la medición del rendimiento. Véase E. D. Pulakos, R. Mueller-Hanson y S. Arad, «The Evolution of Performance Management. Searching for Value», *Annual Review of Organizational Psychology and Organizational Behavior*, 6 (2018), pp. 249-271.

[3] S. E. Scullen, M. K. Mount y M. Goff, «Understanding the Latent Structure of Job Performance Ratings», *Journal of Applied Psychology*, 85 (2000), pp. 956-970.

[4] Un pequeño componente —el 10 por ciento de la varianza total en algunos estudios— es lo que los investigadores llaman la «perspectiva del evaluador» o el «efecto de nivel», en el sentido de nivel en la organización, no del «ruido de nivel» como lo definimos aquí. La perspectiva del evaluador refleja que, al calificar a la misma persona, un jefe difiere sistemáticamente de un compañero, y un compañero de un subordinado. Una interpretación comprensiva de los resultados de los sistemas de calificación de 360 grados podría argumentar que esto no es ruido. Si personas de distintos niveles de la organización ven sistemáticamente diferentes facetas del rendimiento de la misma persona, su juicio sobre esa persona debería diferir sistemáticamente, y sus calificaciones deberían reflejarlo.

[5] S. E. Scullen, M. K. Mount y M. Goff, «Understanding the Latent Structure...»; C. Viswesvaran, D. S. Ones y F. L. Schmidt, «Comparative Analysis of the Reliability of Job Performance Ratings», *Journal of Applied Psychology*, 81 (1996), pp. 557-574;G. J.Greguras y C. Robie, «A New Look at Within-Source Interrater Reliability of 360-Degree Feedback Ratings», *Journal of Applied Psychology*, 83 (1998), pp. 960-968; G. J. Greguras, C. Robie, D. J. Schleicher y M. A. Goff, «A Field Study of the Effects of Rating Purpose on the Quality of Multisource Ratings», *Personnel Psychology*, 56 (2003), pp. 1-21; C. Viswesvaran, F. L. Schmidt y D. S. Ones, «Is There a General Factor in Ratings of Job Performance? A Meta-Analytic Framework for Disentangling Substantive and Error Influences», *Journal of Applied Psychology*, 90 (2005), pp. 108-131; B. Hoffman, C. E. Lance, B. Bynum y W. A. Gentry, «Rater Source Effects Are Alive and Well After All», *Personnel Psychology*, 63 (2010), pp. 119-151.

[6] K. R. Murphy, «Explaining the Weak Relationship Between Job Performance and Ratings of Job Performance», *Industrial and Organizational Psychology*, 1 (2008), pp. 148-160, en particular p. 151.

[7] En el análisis de las fuentes de ruido, ignoramos la posibilidad de que el ruido de los casos se deba a sesgos sistemáticos en la calificación de determinados empleados o categorías de empleados. Ninguno de los estudios que pudimos localizar sobre la variabilidad de las calificaciones de rendimiento las comparó con un rendimiento «verdadero» evaluado externamente.

[8] E. D. Pulakos y R. S. O'Leary, «Why Is Performance Management Broken?», *Industrial and Organizational Psychology*, 4 (2011), pp. 146-164; M. M. Harris, «Rater Motivation in the Performance Appraisal Context. A Theoretical Framework», *Journal of Management*, 20 (1994), pp. 737-756; K. R. Murphy y J. N. Cleveland, *Understanding Performance Appraisal. Social, Organizational, and Goal-Based Perspectives*, Thousand Oaks, California, Sage, 1995.

[9] G. J. Greguras *et al*., «A Field Study...».

[10] P. W. Atkins y R. E. Wood, «Self- Versus Others' Ratings As Predictors of Assessment Center Ratings. Validation Evidence for 360-Degree Feedback Programs», *Personnel Psychology*, 55, 4 (2002), pp. 871-904.

[11] *Ibid*.

[12] Olson y Davis, citado en Peter G. Dominick, «Forced Ranking. Pros, Cons and Practices», en *Performance Management. Putting Research into Action*, James W. Smither y Manuel London (eds.), San Francisco, Jossey-Bass, 2009, pp. 411-443.

[13] P. G. Dominick, «Forced Ranking...».

[14] Barry R. Nathan y Ralph A. Alexander, «A Comparison of Criteria for Test Validation: A Meta-Analytic Investigation», *Personnel Psychology*, 41, 3 (1988), pp. 517-535.

[15] Adaptado de Richard D. Goffin y James M. Olson, «Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others», *Perspectives on Psychological Science*, 6, 1 (2011), pp. 48-60.

[16] M. Buckingham y A. Goodall, «Reinventing Performance Management», *Harvard Business Review*, 1 de abril de 2015, pp. 1-16 (doi: ISSN: 0017-8012).

[17] Corporate Leadership Council, citado en S. Adler *et al*., «Getting Rid of Performance Ratings. Genius or Folly? A Debate», *Industrial and Organizational Psychology*, 9 (2016), pp. 219-252.

[18] E. D. Pulakos, R. Mueller-Hanson y S. Arad, «Evolution of Performance Management...», p. 250.

[19] A. Tavis y P. Cappelli, «The Performance Management Revolution», *Harvard Business Review*, octubre de 2016, pp. 1-17.

[20] Frank J. Landy y James L. Farr, «Performance Rating», *Psychological Bulletin*, 87, 1 (1980), pp. 72-107.

[21] D. J. Woehr y A. I. Huffcutt, «Rater Training for Performance Appraisal.A Quantitative Review», *Journal of Occupational and Organizational Psychology*, 67 (1994), pp. 189-205; S. G. Roch, D. J. Woehr, V. Mishra y U. Kieszczynska, «Rater Training Revisited. An Updated Meta-Analytic Review of Frame-of-Reference Training», *Journal of Occupational and Organizational Psychology*, 85 (2012), pp. 370-395; M. H. Tsai, S. Wee y B. Koh, «Restructured Frame-of-Reference Training Improves Rating Accuracy», *Journal of Organizational Behavior*, (2019), pp. 1-18 (doi:10.1002/job.2368).

[22] El panel izquierdo es una adaptación de Richard Goffin y James M. Olson, «Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others», *Perspectives on Psychological Science*, 6, 1 (2011), pp. 48-60.

[23] S. G. Roch *et al*., «Rater Training Revisited...».

24. Estructura en la contratación

[24] Ernest O'Boyle y Herman Aguinis, «The Best and the Rest. Revisiting the Norm of Normality of Individual Performance», *Personnel Psychology*, 65, 1 (2012), pp. 79-119; Herman Aguinis y Ernest O'Boyle, «Star Performers in Twenty-First Century Organizations», *Personnel Psychology*, 67, 2 (2014), pp. 313-350.

[1] A. I. Huffcutt y S. S. Culbertson, «Interviews», en S. Zedeck (ed.), *APA Handbook of Industrial and Organizational Psychology*, Washington D. C., American Psychological Association, 2010, pp.185-203.

[2] N. R. Kuncel, D. M. Klieger y D. S.Ones, «In hiring, algorithms beat instinct», *Harvard Business Review*, 92, 5 (2014), p. 32.

[3] R. E. Ployhart, N. Schmitt y N. T. Tippins, «Solving the Supreme Problem. 100 Years of Selection and Recruitment at the *Journal of Applied Psychology*», *Journal of Applied Psychology*, 102 (2017), pp. 291-304.

[4] M. McDaniel, D. Whetzel, F. L. Schmidt y S. Maurer, «Meta Analysis of the Validity of Employment Interviews», *Journal of Applied Psychology*, 79 (1994), pp. 599-616; A. Huffcutt y W. Arthur, «Hunter and Hunter (1984) Revisited. Interview Validity for Entry-Level Jobs», *Journal of Applied Psychology*, 79 (1994), p. 2; F. L. Schmidt y J. E. Hunter, «The Validity and Utility of Selection Methods in Personnel Psychology. Practical and Theoretical Implications of 85 Years of Research Findings», *Psychology Bulletin*, 124 (1998), pp. 262-274; F. L. Schmidt y R. D. Zimmerman, «A Counterintuitive Hypothesis About Employment Interview Validity and Some Supporting Evidence», *Journal of Applied Psychology*, 89 (2004), pp. 553-561. Obsérvese que la validez es mayor cuando se consideran ciertos subconjuntos de estudios, en particular si la investigación utiliza evaluaciones del rendimiento creadas específicamente para este fin, en lugar de las evaluaciones administrativas existentes.

[5] S. Highhouse, «Stubborn Reliance on Intuition and Subjectivity in Employee Selection», *Industrial and Organizational Psychology*, 1 (2008), pp. 333-342; D. A. Moore, «How to Improve the Accuracy and Reduce the Cost of Personnel Selection», *California Management Review*, 60 (2017), pp. 8-17.

[6] L. A. Rivera, «Hiring as Cultural Matching. The Case of Elite Professional Service Firms», *American Sociology Review*, 77 (2012), pp. 999-1022.

[7] F. L. Schmidt y R. D. Zimmerman, «A Counterintuitive Hypothesis...»; Timothy A.Judge, Chad A.Higgins y Daniel M. Cable, «The Employment Interview. A Review of Recent Research and Recommendations for Future Research», *Human Resource Management Review*, 10 (2000), pp. 383-406; A. I. Huffcutt, S. S. Culbertson y W. S. Weyhrauch, «Employment Interview Reliability. New Meta-Analytic Estimates by Structure and Format», *International Journal of Selection and Assessment*, 21 (2013), pp. 264-276.

[8] M. R. Barrick *et al*., «Candidate Characteristics Driving Initial Impressions During Rapport Building. Implications for Employment Interview Validity», *Journal of Occupational and Organizational Psychology*, 85 (2012), pp. 330-352; M. R. Barrick, B. W. Swider y G. L. Stewart, «Initial Evaluations in the Interview. Relationships with Subsequent Interviewer Evaluations and Employment Offers», *Journal of Applied Psychology*, 95 (2010), p. 1163.

[9] G. L. Stewart, S. L. Dustin, M. R. Barrick y T. C. Darnold, «Exploring the Handshake in Employment Interviews», *Journal of Applied Psychology*, 93 (2008), pp. 1139-1146.

[10] T. W. Dougherty, D. B. Turban y J. C. Callender, «Confirming First Impressions in the Employment Interview. A Field Study of Interviewer Behavior», *Journal of Applied Psychology*, 79

(1994), pp. 659-665.

[11] J. Dana, R. Dawes y N. Peterson, «Belief in the Unstructured Interview. The Persistence of an Illusion», *Judgment and Decision Making*, 8 (2013), pp. 512-520.

[12] Nathan R. Kuncel *et al*., «Mechanical versus Clinical Data Combination in Selection and Admissions Decisions. A Meta-Analysis», *Journal of Applied Psychology*, 98 (2013), pp. 1060-1072.

[13] Laszlo Bock, entrevista con Adam Bryant, *The New York Times*, 19 de junio de 2013. Véase también Laszlo Bock, *Work Rules! Insights from Inside Google That Will Transform How You Live and Lead*, Nueva York, Hachette, 2015.

[14] C. Fernández-Aráoz, «Hiring Without Firing», *Harvard Business Review*, 1 de julio de 1999.

[15] Para una guía accesible de las entrevistas estructuradas, véase Michael A. Campion, David K. Palmer y James E. Campion, «Structuring Employment Interviews to Improve Reliability, Validity and Users' Reactions», *Current Directions in Psychological Science*, 7, 3 (1998), pp. 77-82.

[16] J. Levashina, C. J. Hartwell, F. P. Morgeson y M. A. Campion, «The Structured Employment Interview. Narrative and Quantitative Review of the Research Literature», *Personnel Psychology*, 67 (2014), pp. 241-293.

[17] M. McDaniel *et al*., «Meta Analysis...»; A. Huffcutt y W. Arthur, «Hunter and Hunter (1984) Revisited...»; F. L. Schmidt y J. E. Hunter, «Validity and Utility...»; F. L. Schmidt y R. D. Zimmerman, «A Counterintuitive Hypothesis...».

[18] F. L. Schmidt y J. E. Hunter, «Validity and Utility...».

[19] Daniel Kahneman, *Thinking, Fast and Slow*, Nueva York, Farrar, Straus and Giroux, 2011, p. 229. (Hay trad. cast.:*Pensar rápido, pensar despacio*, Barcelona, Debate, 2012.)

[20] N. R. Kuncel, D. M. Klieger y D. S. Ones, «In Hiring, Algorithms Beat Instinct». Véase también M. A. Campion, D. K. Palmer y J. E. Campion, «Structuring Employment Interviews...».

25. El protocolo de evaluaciones mediadoras

[21] J. Dana, R. Dawes y N. Peterson, «Belief in the Unstructured Interview...».

[1] Daniel Kahneman, Dan Lovallo y Olivier Sibony, «A Structured Approach to Strategic Decisions. Reducing Errors in Judgment Requires a Disciplined Process», *MIT Sloan Management Review*, 60 (2019), pp. 67-73.

[2] Andrew H. Van de Ven y André Delbecq, «The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes», *Academy of Management Journal*, 17, 4 (1974), pp. 605-621. Véase también el capítulo 21.

26. Los costes de reducir el ruido

[1] Kate Stith y José A. Cabranes, *Fear of Judging. Sentencing Guidelines in the Federal Courts*, Chicago, University of Chicago Press, 1998, p. 177.

[1] Albert O. Hirschman, *The Rhetoric of Reaction. Perversity, Futility, Jeopardy*, Cambridge, Massachusetts, Belknap Press, 1991.

[2] K. Stith y J. Cabranes, *Fear of Judging*...

[3] Véase, por ejemplo, «Three Strikes Basics», Stanford Law School, <https://law.stanford.edu/stanford-justice-advocacy-project/three-stri kes-basics/>.

[4] 428 U. S. 280 (1976).

[5] Cathy O'Neil, *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*, Nueva York, Crown, 2016.

[6] Will Knight, «Biased Algorithms Are Everywhere, and No One Seems to Care», *MIT Technology Review*, 12 de julio de 2017.

[7] JeffLarson, Surya Mattu, Lauren Kirchner y Julia Angwin, «How We Analyzed the COMPAS Recidivism Algorithm», *ProPublica*, 23 de mayo de 2016, <www.propublica.org/article/how-we-analyzed-the-compas-re cidivism-algorithm>. La afirmación de presencia de sesgo en este ejemplo es discutida, y diferentes definiciones de sesgo pueden llevar a conclusiones opuestas. Para conocer las opiniones sobre este caso y, en general, sobre la definición y medición del sesgo algorítmico, véase la nota 10 de este capítulo 26.

[8] Aaron Shapiro, «Reform Predictive Policing», *Nature*, 541, 7638 (2017), pp. 458-460.

[9] Aunque esta preocupación resurge en el contexto de los modelos basados en la inteligencia artificial (IA), no es específica de esta. Ya en 1972, Paul Slovic señaló que modelizar la intuición preservaría y reforzaría, y quizá incluso magnificaría, los sesgos cognitivos existentes. Paul Slovic, «Psychological Study of Human Judgment. Implications for Investment Decision Making», *Journal of Finance*, 27 (1972), p. 779.

27. Dignidad

[10] Para una introducción a este debate en el contexto de la controversia sobre el algoritmo de predicción de la reincidencia COMPAS, véase J. Larson *et al*., «How We Analyzed the COMPAS Recidivism Algorith»; William Dieterich *et al*., «COMPAS Risk Scales. Demonstrating Accuracy Equity and Predictive Parity», Northpointe, 8 de julio de 2016, <http:// go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commen tary_Final_070616.pdf>; Julia Dressel y Hany Farid, «The Accuracy, Fairness, and Limits of Predicting Recidivism», *Science Advances*, 4, 1 (2018), pp. 1-6; Sam Corbett-Davies *et al*., «A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear», *The Washington Post*, 17 de octubre de 2016, <www.washing tonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithmbe-racist-our-analysis-is-more-cautious-than-propublicas>; Alexandra Chouldechova, «Fair Prediction with Disparate Impact. A Study of Bias in Recidivism Prediction Instruments», *Big Data*, 153 (2017), p. 5; Jon Kleinberg, Sendhil Mullainathan y Manish Raghavan, «Inherent Trade-Offs in the Fair Determination of Risk Scores», Leibniz International Proceedings in Informatics (enero de 2017).

[1] Tom R. Tyler, *Why People Obey the Law*, 2.ª ed., New Haven, Connecticut, Yale University Press, 2020.

[2] Cleveland Board of Education *vs*. LaFleur, 414 U. S. 632 (1974).

[3] Laurence H. Tribe, «Structural Due Process», *Harvard Civil RightsCivil Liberties Law Review*, 10, 2 (primavera de 1975), p. 269.

[4] K. Stith y J. Cabranes, *Fear of Judging*..., p. 177.

28. ¿Normas o estándares?

[5] Véase, por ejemplo, Philip K. Howard, *The Death of Common Sense. How Law Is Suffocating America*, Nueva York, Random House, 1995, y Philip K. Howard, *Try Common Sense. Replacing the Failed Ideologies of Right and Left*, Nueva York, W. W. Norton & Company, 2019.

[1] «Hate Speech», Facebook. Community Standards, <www.facebook.com/communitystandards/hate_speech>.

[2] Andrew Marantz, «Why Facebook Can't Fix Itself», *The New Yorker*, 12 de octubre de 2020.

[3] Jerry L. Mashaw, *Bureaucratic Justice*, New Haven, Connecticut, Yale University Press, 1983.

«Para viajar lejos no hay mejor nave que un libro.»

EMILY DICKINSON

Gracias por tu lectura de este libro.

En **Penguinlibros.club** encontrarás las mejores recomendaciones de lectura.

Únete a nuestra comunidad y viaja con nosotros.

Penguinlibros.club

Penguin Random House Grupo Editorial

Penguinlibros

Índice

Ruido. Un fallo en el juicio humano

Introducción. Dos clases de error
Primera parte. Buscando el ruido
- 1. Crimen y castigo ruidoso
- 2. Un sistema ruidoso
- 3. Decisiones singulares

Segunda parte. Nuestra mente es un instrumento de medición
- 4. Cuestiones de juicio
- 5. La medición del error
- 6. El análisis del ruido
- 7. Ruido de ocasión
- 8. Cómo los grupos amplifican el ruido

Tercera parte. El ruido en los juicios predictivos
- 9. Juicios y modelos
- 10. Reglas sin ruido
- 11. Ignorancia objetiva
- 12. El valle de lo normal

cuarta parte. Cómo se genera el ruido
- 13. Heurísticas, sesgos y ruido
- 14. La operación de emparejamiento
- 15. Escalas
- 16. Patrones
- 17. Las fuentes de ruido

Quinta parte. Mejorar los juicios
- 18. Mejores juzgadores para mejores juicios
- 19. Supresión de sesgos e higiene de las decisiones
- 20. Secuenciación de la información en la ciencia forense
- 21. Selección y suma en los pronósticos
- 22. Directrices en la medicina
- 23. Definición de la escala en las calificaciones
- 24. Estructura en la contratación
- 25. El protocolo de evaluaciones mediadoras

Sexta parte. Ruido óptimo
- 26. Los costes de reducir el ruido
- 27. Dignidad
- 28. ¿Normas o estándares?

Resumen y conclusión. Tomar el ruido en serio

Epílogo. Un mundo menos ruidoso

Agradecimientos

Apéndice A. Cómo llevar a cabo una auditoría del ruido

Apéndice B. Una lista de control para un observador

Apéndice C. Corrección de las predicciones

Sobre este libro

Sobre los autores

Créditos

Notas

This file was downloaded from Z-Library project

Your gateway to knowledge and culture. Accessible for everyone.

z-library.sk z-lib.gs z-lib.fm go-to-library.sk

Official Telegram channel

Z-Access

https://wikipedia.org/wiki/Z-Library